Schema mode assessment through a conversational agent

David Allaart
Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology
Email: D.G.Allaart@student.tudelft.nl

***Abstract*—Schema therapy is a psychotherapy for treatment of personality disorders and other psychological disorders. An important element of schema therapy is determining a patient's schema modes, a concept central to deciding the treatment approach. The objective of this work was to evaluate whether assessment of these schema modes is possible through a conversational agent. The agent designed held an interview style conversation, where first an open story was requested, which was then automatically analysed and followed up with evaluating questions. The results indicated the outcome of the agent was a significant predictor for a person's schema modes, in that a schema mode confirmed by the agent was 5.20 times more likely to be confirmed through standard means than a schema mode that was not confirmed by the agent. The results also hinted at a non-inferior user experience and time savings when using the agent instead of the questionnaire.**

***Index Terms*—schema therapy; mental health; conversational agent; NLP; adaptive questionnaire**

I. Introduction

A. Motivation

Mental illness accounts for about one-third of the world's disability caused by adult health problems [1]. In the USA, 46% of people qualify for a mental illness at some point in their life [2], and several community samples seem to indicate that around 1 in 7 adults deal with personality disorders (PD) [3], [4]. PD's in particular are associated with a severe reduced quality of life [5] as well as high societal costs [6]. Additionally, programs designed to treat PD's are scarce and hard to access, and health professionals often lack training in the treatment of these disorders [7].

One of the types of therapy aimed at fixing this problem is schema therapy. Introduced by Young in 2003 [8], schema therapy is a treatment for complex psychological problems. It is a popular way of treating personality disorders and chronic DSM Axis I disorders, especially when patients do not respond well to traditional treatment or relapse to old behaviour [9]. A central construct in schema therapy is the idea of schema modes, which are momentary mind states that every person experiences now and then. In psychologically stable people, these states are flexible and mild, but in people with personality disorders, they can be strong and rigid, and can seem like completely separate parts of someone's personality. Assessing a person's schema modes is an important part of schema therapy, as understanding a person's schema modes is essential for a therapist to provide appropriate therapy. This is currently done with the Short Schema Mode Index (SMI) [10], a 118 item questionnaire which is scored using a 6-point scale ranging from "never or hardly ever" to "always". The 118 items on the questionnaire relate to 1 of 14 schema modes.

However, there are limitations to this. The questionnaire is long, and takes approximately 40 minutes to fill in [10]. Additionally, after the questionnaire is taken, the therapist discusses the results with the patient to develop a mode model for the patient. It takes about 3-6 sessions to establish this model [11], [12], and touches on situations and thought patterns in daily life as well as the past. While schema therapy is a cost-effective treatment [13], there is still a significant societal need for affordable mental healthcare and a general lack of resources to achieve this [14], meaning that a reduction in the amount of sessions needed to establish a mode model would make schema therapy even more interesting.

Furthermore, there is the problem that schema modes are not constant, and are momentary states, that can "flip" due to life events or moods [8], [15], [16], [17]. Because of the length of the SMI questionnaire, it is generally taken only once. Consequently, the SMI does not measure these schema mode flips, resulting in only a single static measurement of what actually is a dynamic system of schema modes.

Our vision for this project was to automate a part of the assessment process with a conversational agent. A conversational agent (occasionally known as a "chatbot") is a computer system intended to converse with a human being. While the concept of a conversational agent has been around since the 1960's [18], there has been a lot of development in recent years, and conversational agents are now in widespread use in entertainment, business and healthcare [19]. Conversational agents are often used in mental healthcare [20], [21], both for assessment [22], [23] as well as interventions [24], [25]. The questions of the SMI are a good fit for being adapted to a conversational agent, as answering them is already done unsupervised, and research shows that there is no significant difference in the way people answer psychological questionnaires when being interviewed by a conversational agent compared to filling it in unsupervised [26], [27]. Unlike a questionnaire, a conversational agent can be flexible, and can ask questions related to an event as it is happening, reducing the need to fill in questions irrelevant to the situation. The 14-factor nature of the SMI allows for splitting the questionnaire on the suspected relevant schema modes. Furthermore, while a questionnaire is static and requires all questions to be filled in, a conversational agent could use existing norms of clinical relevance to determine whether further questions need

to be asked, or whether the existing answers are sufficient, further reducing the amount of questions that need to be asked. Questions are also easily randomized, which reduces the learning effect associated with repeatedly filling in the same questionnaire.

On top of the advantages a conversational agent can deliver for the taking of the SMI, it can go beyond just asking static questions. A conversational agent can also be used to analyse qualitative data, such as a patient's recount of a recent emotional story. Going over and analysing emotional events and a patient's response to those is a part of the 3-6 sessions following the taking of the SMI. A conversational agent that is able to analyse these situations outside of patient-therapist contact hours, and can do this independently and automatically, could be a separate measure in and of itself. This can be used to provide valuable information to a therapist, which can save them time.

Because of this, we suspected that a conversational agent could be a powerful tool in the assessment of schema modes, allowing for multiple Ecological Momentary Assessments (EMA), which can give the therapist more information than a single questionnaire, potentially reducing the amount of sessions needed to come to a schema mode model for the patient. Finally, since the data acquired by the conversational agent will be digital and easier to analyse, one could foresee that data bringing new insights for schema therapists regarding their patients, or perhaps for schema therapy as a whole.

B. Research Question

The main question this research has answered is as follows: **How can a conversational agent be used to assess a person's schema modes?**

This was then further distilled into the following sub-questions:

- What should the design requirements be for a conversational agent to be able to assess a person's schema modes?
- How should these design requirements be implemented to realise a conversational agent that can assess a person's schema modes?
- How well can a conversational agent be used to assess a person's schema modes?

C. Approach

To come up with the design requirements, we evaluated related work and consulted with experts in the field. This culminated in a set of design requirements and methods to establish the quality of the conversational agent once realized. The results of this can be found in Section II. Based on these design requirements, we set our design specifications and built a conversational agent, which was described in Section III. We then designed an observational experiment to evaluate the quality of the conversational agent. Participants interacted with the conversational agent and were asked to retell a recent emotional event, which the conversational agent analysed. Based on the analysis of this story, the conversational agent asked questions from the SMI to evaluate whether the preliminary result was accurate. The perceived usability and the time taken by participants was also evaluated. The results of this experiment can be found in Section IV. Finally, we drew conclusions from this, as well as some suggestions for future work. This can be found in Section V.

II. Related Work

This section discusses the requirements of the conversational agent. What does it need to be able to do to assess a person's schema modes? First, relevant examples from literature are discussed. To ensure a proper foundation of the conversational agent in the target domain, those findings were evaluated with an expert panel. Finally, the requirement specification is detailed at the end of this section.

A. Literature review

Fortunately, this project is not the first to use a conversational agent for mental healthcare delivery. Already in 1966, one of the first conversational agents was suggested to be used for psychotherapy [18]. More recently, there have been success stories such as Woebot [25], which boasts over 4.7 million messages exchanged every week, and 75% users feeling better after using it just once[1]. While not all available applications are a product of academic research, plenty are published, and there are also some review papers on the topic.

One of these papers is a literature study of both research and developments of mental health apps by Bakker et al [21]. They evaluated several mental health apps, and formulated 16 requirements for future developments of mental health applications. All requirements are rated on the strength of their associated evidence, and guidelines are given on how to implement these requirements. Some are more standard, such as to have a simple and logical interface, or to encourage users to be open and honest when self-reporting, but some are more specific. One recommendation is to have a unique conversation, that is to say that what the conversational agent says should not be static, but adapted to user input. Furthermore, they recommend to have the assessment as close to the triggering emotional event, when the event is still fresh in memory, and to explicitly design the system in such a way that daily assessment is possible.

Another design challenge is the use of open text. A result of allowing users to write open text is that the variety in potential user responses increases. Elmasri and Maeder recently developed a conversational agent [24] for mental health interventions in young adults (18-25 years old) to assess alcohol drinking habits. One of the things they note in their research is that users were critical about the conversational agent only recognizing a limited amount of keywords, and they recommended to make sure that the conversational agent recognizes a wide range of keywords.

Furthermore, since this piece of software was made for the mental health domain, ease of use and availability of information from the development side was even more important than usual. Denecke et al [22] note that especially for (mental) health applications, not all the medical staff that interacts with

[1] https://woebothealth.com/, consulted 29-11-2020

the system has the required technical background to operate a complicated system. Things like keyword lists used for automatic classification and response formulations will likely change over time, and it should be as easy as possible for experts in the mental health domain to perform this kind of maintenance or iterative improvement.

B. Expert Panel

One common thread that was mentioned in many of the reviewed papers was the importance of integrating the solution in the existing mental health care delivery process [20], [24], [22]. This is important to reach the public health potential of the technology. Mohr [20] specifically mentions that technology such as this "will have to be accepted and adopted by healthcare delivery teams as well as patients."

To ensure this, we contacted two schema mode therapy specialists to form an expert panel. The experts were enthusiastic about the potential for EMA's, and noted the value of new measure for assessing schema modes, besides the SMI that is used currently. While our experts agreed with the requirements mentioned in subsection II-A, they disagreed with some of the other recommendations in those publications. While multiple authors [22], [24] suggest having an empathetic, understanding tone in the conversation, the experts instead recommended a neutral tone. The reason behind this is that in individuals with schema problems, empathetic feedback may create a mode flip [8]. Additionally, the results of the assessment should not immediately be shared with the user, as this may also induce similar flips.

While the experts agreed that the application needed to be integrated into the existing mental health delivery process, they also stressed that it should be an automatic and independent system. This means a "hands-off" system, which should not require real-time interaction or monitoring from therapists while the patient is interacting with the system. While maintenance is always required, requiring a therapist to run the system in real-time would not result in any benefit for the therapist, time-wise.

C. Requirements

From the feedback of the expert panel and the literature a requirement list was specified, which can be found in Table I in no particular order.

1	The conversational agent should be a new measure, independent of SMI.
2	The conversational agent should have the option for open questions, not just closed questions.
3	The conversational agent should have a simple and logical interface.
4	The conversational agent should encourage users to be open about self-reporting.
5	The response of the conversational agent should be unique and adapted to user input.
6	Maintenance on the conversational agent should require as little technical knowledge as possible.
7	If the conversational agent uses training data or something similar for automatic classification, this dataset should be representative and comprehensive.
8	The conversational agent should be an independent system that does not require active operation by a therapist.

TABLE I: List of requirements compiled from the feedback of the literature and the expert panel.

III. System Design

In this section, we will specify the design requirements, and discuss how we subsequently implemented these specifications in our conversational agent.

A. Conversation design

The structure of the conversation with a user can be found in Figure 1. During the conversation the agent follows the standard structural schema composed by Robinson [28]. The first stage is the establishing of the reason for the encounter, where the agent asks the user to tell a recent emotional event. The user could respond with open text. The conversational agent then analyses the story of the user, resulting in a set of schema modes of which the agent thinks are relevant to the story. The agent then moves to the second stage, where it gathers additional information to refine the assessment. The agent does this by asking the user questions from the SMI questionnaire that correspond to the schema modes it is investigating. Once the agent receives enough answers to draw its result, it thanks the user for their time and ends the conversation. The agent does not share the final assessment of the user's schema modes with the user.

B. Adaptive questionnaire

To reduce the amount of questions, the conversational agent uses an adaptive questionnaire strategy [29] when asking the follow-up questions. This is one of the main advantages of using a conversational agent for assessment. Where a pen-and-paper questionnaire requires all questions to be filled in, a computerized system can notice that a user has already filled in enough questions to pass the threshold of (clinical) relevance, and save the user time by skipping the rest of the questions.

Since the original SMI questionnaire is also intended to have its questions answered independently of each other, the order in which to ask the questions is irrelevant. The text analysis algorithm provides the agent with a pool of potential questions, and whenever the agent wants to ask another question, it randomly selects one of the questions from that pool.

The remaining issue is when the conversational agent can stop asking questions. While the conversation obviously ends when the pool of questions is empty, the goal of using an adaptive questionnaire is to allow users to answer fewer questions. To facilitate this, we need to establish when the user has answered questions related to a particular schema mode positively enough that the conversational agent does not need to continue asking questions.

For this, we turned to the official SMI form[2]. The questions are answered on a scale from 1-6, and a schema mode is considered confirmed if any item related to that schema mode is answered with a 5 or a 6, or if the average answer of the items related to that schema mode is at least 3.5. The

[2]https://www.schematherapie.nl/document/Schemamodi-vragenlijst-SMI-Engels.xlsx, visited 02-02-2021

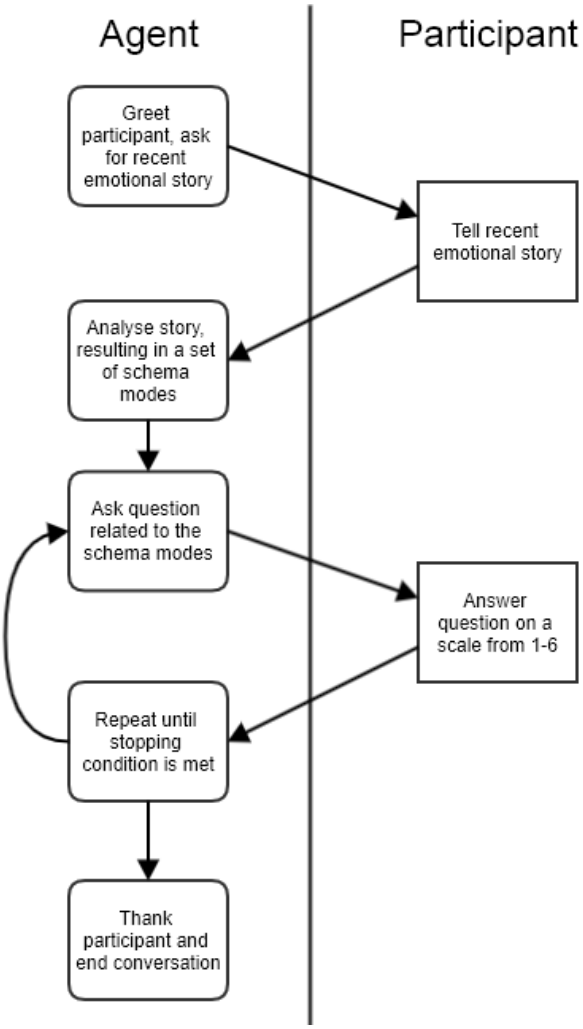

Fig. 1: Visual representation of conversation design with the conversational agent.

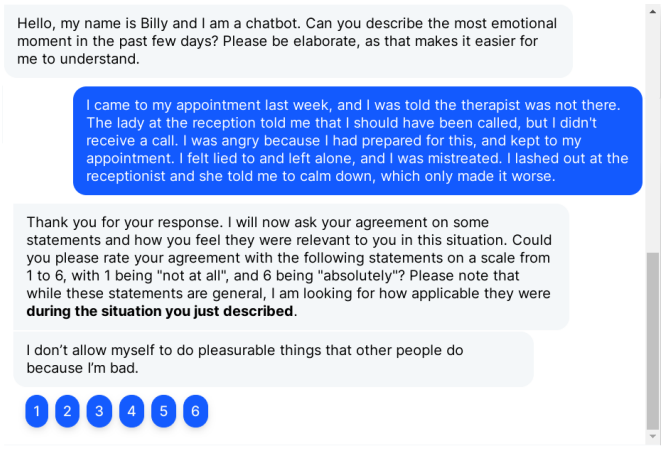

Fig. 2: Interface of the conversational agent implemented with Rasa Webchat. The story is an artificial example created by the authors.

conversational agent also applies this logic, with the one caveat that it also has to take into account the unanswered questions for calculating the average score. As such, it calculates the mean of the answers under the assumption that the unanswered questions are given a 1, so that the final criteria is only met when the mean of all answers cannot be lower than 3.5. If either of these conditions occur for a particular schema mode, the agent considers that schema mode completed, and all questions relating to that schema mode are removed from the pool of remaining questions.

C. Rasa

To implement the conversational agent we employed Rasa[3], an infrastructure platform for making conversational agents. It is built on Rasa Open Source and has a built-in Natural Language Understanding (NLU) capability, which we used for the automatic classification of the recent emotional story provided by the user. The full configuration of the conversational agent can be found in Appendix D. The text analysis algorithm evaluates the story and ranks the schema modes on relevance to the story. This ranking is then used by the agent to select questions to ask the user.

Another important feature of Rasa is that there is a clear separation of content and application logic. Classification in Rasa works with so-called "intents", which are defined by providing a sample of words and phrases that would count as instances of this intent. A full list of the intents used for this conversational agent can be found in Appendix E. It is important to note that the application logic would not be affected by changes in these lists of words and phrases. This means that the agent can easily be improved in an iterative manner, by making changes to the aforementioned lists of words and phrases.

Finally, the interface was implemented with Rasa Webchat[4], resulting in the interface that can be seen in Figure 2.

D. Training data

Since there was no standard corpus available to train the text analysis algorithm on, the lists of words and phrases related to the different schema modes were drafted by the authors to train the conversational agent. This list was compiled based on many different sources, such as the original SMI questionnaire, therapist training DVD's[5] as well as literature research [30], [11], [8]. At the recommendation of the expert panel, only 7 out of the 14 schema modes were used in this research, to note: Angry Child, Detached Protector, Happy Child, Healthy Adult, Impulsive Child, Punishing Parent and Vulnerable Child. The intent lists related to these schema modes were also reviewed by the expert panel, and can be found in Appendix E.

E. Verification

To verify that the implementation of the conversational agent could recognize the schema modes through the intent lists, a small set of example stories was produced by the

[3]https://rasa.com/

[4]https://github.com/botfront/rasa-webchat

[5]https://www.schematherapie.nl/vakinformatie/dvds/

authors. These were also based on the study of the aforementioned literature and training DVD's, and were written to be a vignette of a particular schema mode. These were also submitted to the expert panel, who were asked to select the corresponding schema mode. The resulting data was compared to the modes the vignettes were supposed to represent, which was considered the ground truth. To measure inter-rater reliability, Cohen's Kappa was calculated. As can be seen in Table II, the average value of the kappa for the agreement of the experts in the expert panel with the ground truth was 0.66. The kappa for the agreement between the experts themselves was 0.65.

	Ground truth	Rater 1	Rater 2
Ground truth			
Rater 1	0.75		
Rater 2	0.58	0.65	
Average	0.66	0.65	

TABLE II: Inter-rater reliability between the ground truth and the different raters in the expert panel. Listed values are Cohen's Kappa values.

Additionally, these vignettes were evaluated using the text analysis algorithm of the conversational agent. Unfortunately, there were some changes in the Rasa platform between this evaluation and the running of the experiment. This resulted in a forced configuration change in the text analysis algorithm, leading to different results. This was only discovered after running the experiment, at which time the evaluation was ran again. The results of both the before and after can be found in Table III. As can be seen there, the difference in average confidence ratings before the configuration change is smaller than after the configuration change when presented with vignettes designed to evoke that schema mode. This resulted in that the happy and vulnerable schema modes were almost never in the text analysis result. For more details regarding this issue, see Appendix B.

Schema mode	Vignette confidence before	Vignette confidence after
Angry	0.49	0.14
Detached	0.17	0.53
Happy	0.15	0.08
Healthy	0.12	0.43
Impulsive	0.16	0.86
Punishing	0.32	0.80
Vulnerable	0.28	0.02

TABLE III: Vignette evaluation results before and after the configuration change. The confidence is the average confidence score on a scale from 0 (does not fit at all) to 1 (fits perfectly), assigned by the text analysis algorithm to vignettes that were written specifically to invoke that schema mode.

IV. Evaluation

A. Hypotheses

To determine the quality of the conversational agent, we wanted to know how well the conversational agent performed compared to the standard method of assessing schema modes (the SMI questionnaire). Additionally, we wanted to determine the quality of the text analysis algorithm. To evaluate this, we formulated two sub-hypotheses, to note:

> H1a: Compared to a randomly selected schema mode, a schema mode identified by the text analysis process is a better predictor for schema modes being confirmed by the related questions from the SMI about the recent emotional story.

> H1b: Compared to an unconfirmed schema mode, a schema mode confirmed by the related questions from the SMI about the recent emotional story is a better predictor for schema modes confirmed by the SMI questions related to that schema mode.

In the following section, we will use "internal performance" to refer to the concept of H1a, and "external performance" to refer to H1b.

Furthermore, while user experience alone must not be mistaken for efficacy [21], [20], [25], it is an aspect of the overall efficacy of a tool. Research shows that people prefer conversational agents over human interviewers when talking about highly sensitive topics that are likely to evoke negative self-admissions [31], that they prefer interactive over static tasks [32] and perform better in terms of accuracy and speed [33]. As a result of this, we formulated 2 more hypotheses, which are:

> H2: Users experience the usability of interacting with the conversational agent as non-inferior to the process of filling in the SMI.

> H3: A conversational agent can assess schema modes faster than the SMI.

For more background regarding why these hypotheses were selected, please see Appendix A.

B. Study design

The study was designed to be an observational study, to evaluate the hypotheses listed above. During the study, participants were asked to interact with the conversational agent, which requested the participant to write down a story about a recent emotional event they experienced. The agent then analysed this story, and selected the two schema modes that it thought fit best with the story. For evaluating the quality of the text analysis algorithm, the agent then selected one of the remaining 5 schema modes at random and added it to the set of schema modes. The agent then evaluated this set of 3 schema modes with the questions from the SMI related to those schema modes. Separately from the conversational agent interface, the participant also filled in the full SMI. The time taken for both these tasks was measured, and participants also filled in usability questionnaires after each task.

This study received ethical approval from the TU Delft University Human Research Ethics Committee[6]. Before starting

[6]The ID for this research submission is 1257.

data collection, the experimental setup was preregistered with the Open Science Framework [34].

C. Measures

To determine the internal performance of the conversational agent, we scored whether a schema mode was in the top 2 predicted schema modes, or whether it was added as a random third, with a 0 meaning it was random, and a 1 meaning it was top 2. This was then compared to how the participant scored on the conversational agent's questions related to that schema mode. These were rated on a scale from 1-6. As described in Section III, the conversational agent considered a schema mode confirmed if the average score of the questions was 3.5 or higher, or if the participant assigned a 5 or 6 to any of the questions. If a schema mode was confirmed, this variable was a 1, and if not, it was a 0.

For comparing the performance of the agent to the current standard, the participant also took the SMI questionnaire outside of the interface of the conversational agent. From the SMI, 67 questions were selected, corresponding to the 7 modes that the conversational agent was also evaluating and were presented to the participant in a random order. Like the questions that were asked by the conversational agent, these were also answered on a scale from 1-6. From this the "confirmed" measure was calculated in the same way as the conversational agent did. The main measures regarding both internal and external performance are graphically represented in Figure 3.

To measure the usability of the conversational agent, the System Usability Score (SUS) [35] was used. As the SUS is not well suited for absolute information [36], the SUS questionnaire was administered both after the interaction with the conversational agent, as well as after the filling in of the SMI.

Finally, to evaluate how long the participants took, the time was recorded at 4 points: before and after the interaction with the conversational agent as well as before and after the filling in of the SMI. From this the time spent on either task was calculated, which was rounded to the nearest second.

D. Procedure

Data was collected between October 29th and December 1st, 2020. The experiment consisted of 2 parts. In Part 1, the participant interacted with the conversational agent. Before the interaction with the conversational agent, an instructional video was provided on how to interact with the conversational agent. The transcript of this video was also available as plain text. After the participant received the instructions, they were asked two multiple choice questions about the instructions, which functioned as an attention check. If a participant gave the wrong answer on either of the questions, they were excluded from the sample. The participant then interacted with the conversational agent as described in Figure 1. After the interaction concluded, the participant was asked to fill in the SUS questionnaire about their interaction with the conversational agent.

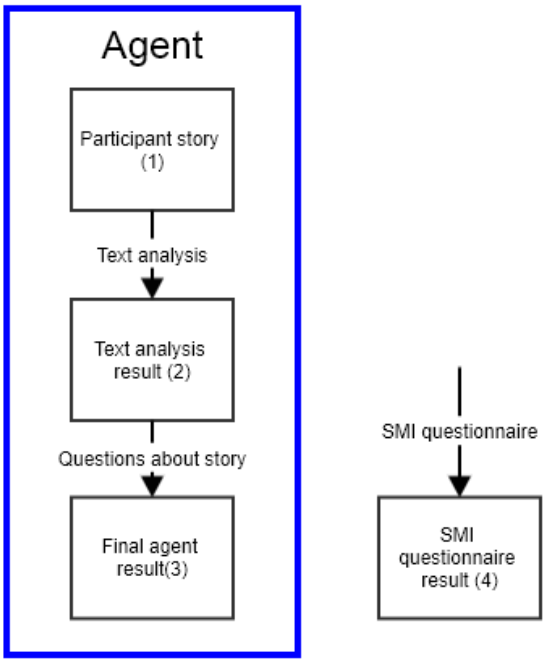

Fig. 3: Schematic representation of the variables regarding performance. Everything in the blue box is happening during the interaction with the agent. For internal performance, the text analysis result (2) is compared to the final agent result (3), and for the external performance, the final agent result is compared to the SMI questionnaire result (4).

Part 2 started with a short instruction for the SMI questionnaire which noted that these questions were about the participant's general life experience, since the questions asked during the conversational agent were regarding the single event they told a story about. They then answered all 67 questions of the SMI, after which the participants were asked fill in the SUS questionnaire about their experience filling in the SMI.

After both Part 1 and Part 2 were concluded, the participant received the debrief message, as well as an opportunity for comments or suggestions. To avoid order effects, participants had a 50/50 chance to do either Part 1 followed by Part 2, or Part 2 followed by Part 1. The full diagram of the experiment can be found in Figure 4.

E. Data preparation

Exclusion criteria for this study were failing the attention checks, or providing the conversational agent with a very short (less than 100 characters) submission instead of an actual story during the interaction with the conversational agent. Participants with incomplete or missing submissions were also excluded.

In total, 933 participants took part in the study. From this, 399 submissions were incomplete. From the 534 submissions that remained, 83 participants failed one or both attention checks. From those who passed the attention checks, 133 submissions had stories that were shorter than 100 characters. Of the remainder, 25 submissions were corrupted because of a server error, leading to a final count of 293 proper submissions.

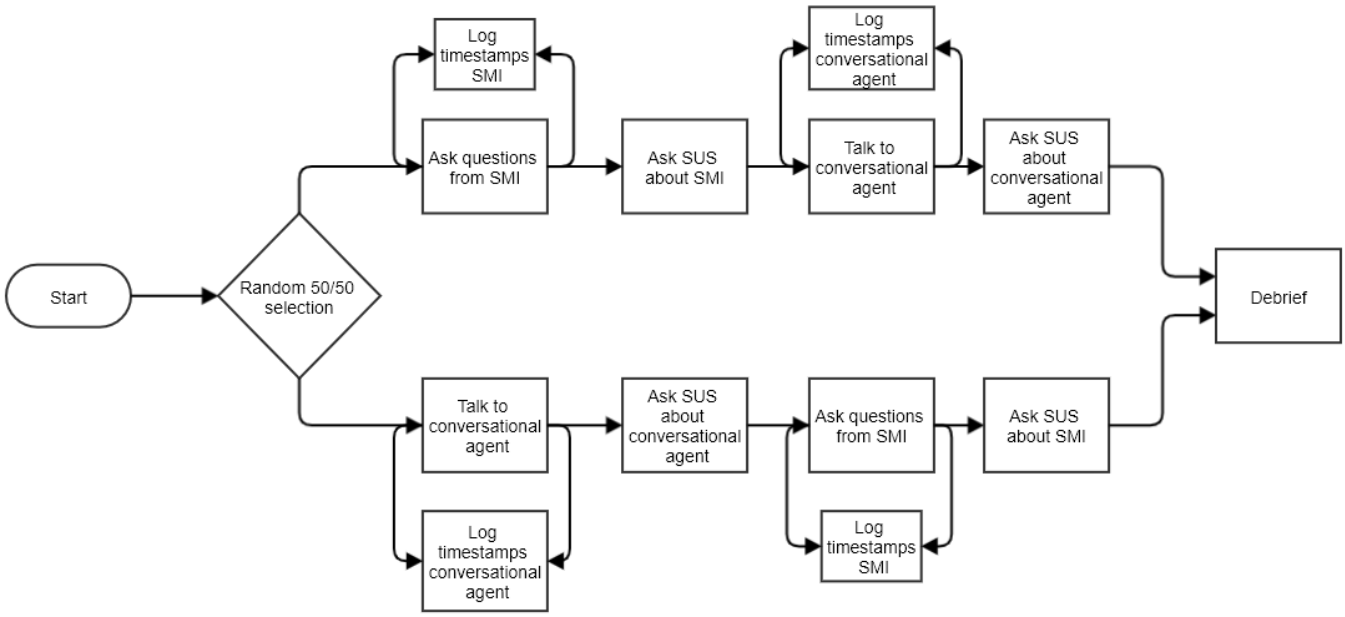

Fig. 4: Schematic overview of the experiment flow.

F. Participants

Data was analysed from 293 English-speaking participants. Participants were recruited through Prolific[7], an online participant recruitment platform. Average completion time was estimated to be 20 minutes[8], for which the participants received a financial compensation of 2 British Pounds. Since the participants were required to perform tasks that depended on writing skills, only native English speakers were selected. Prolific allows for custom pre-screening options, and the ones that were used to achieve this were: "First language" = "English", "Fluent languages" = "English", and "Language Disorders" = "none" or "not applicable".

Mean age of the sample was 32.4 years (SD = 11.1 years, range = 18-81). The age of 1 participant was unavailable. 52.6% of the participants identified as female, 47.1% as male, and $<$0.5% preferred not to say. 71.3% had a UK nationality, 16.7% from the US, and the remaining 12% were nationals of 15 different countries. 50.5% were full-time employed, 15.7% were part-time employed, 13% were currently unemployed but job seeking, 10.6% were not in paid work (e.g. homemaker, retired or disabled), and for the final 10.2% of participants this data was 'other' or unavailable.

G. Analysis

We evaluated the internal performance (H1a) with multi-level analyses. The outcome variable was a binary representation of the final agent result for a particular schema mode, with 0 being that a schema mode was not in the result, and 1 being that it was. Models were built to explore the effects of the text analysis result, the schema mode, as well as the interaction effect of the text analysis result and the schema mode. An overview of the models can be found in Table IV. Model A0 is the basic model and includes only the participants as a random intercept. Model A1 adds the effect of the text analysis result to model A0. Model A2 was built on model A0, and adds the influence of the schema mode. Model A3 is the combination of model A1 and model A2, and so contains the effect of the text analysis result, the effect of the schema mode, and the participants as a random intercept. Finally, model A4 adds the interaction effect to model A3. Due to the issues described in Appendix B, there were not enough measurements for the vulnerable and happy schema modes, and as a result they were dropped from the analysis for internal performance.

Evaluating the external performance (H1b) was also done with multi-level analyses, except here the outcome variable was the result of the SMI questionnaire for a particular schema mode. If a schema mode was not marked as present in the final results of the SMI the outcome variable was 0, and if it was the outcome variable was 1. Again models were built to explore the effects of the final agent result, the schema mode, and the interaction effect between the final agent result and the schema mode. Model B0 is again the basic model and only contains the participants as a random intercept. Model B1 adds the effect of the final agent result, and model B2 adds the influence of the schema mode to model B0. Model B3 again combines model B1 and B2 to include both the final agent result and the schema mode as influences, and model B4 adds the interaction effect to model B3.

We evaluated whether adding the predictors mentioned increased a model's ability to fit the data with ANOVA tests between models and their relative null models. Additionally, we inspected the coefficients of models A4 and B4 to see what the influence of these factors was.

The usability of the conversational agent was evaluated with a non-inferiority test [37]. This type of test requires

[7]https://www.prolific.co/

[8]The actual average time in the sample was 19 minutes and 48 seconds.

	Outcome Variable	Predictors	Interaction effect	Random Intercept
Model A0	Final Agent Result	None	None	Participants
Model A1		Text Analysis Result		
Model A2		Schema Mode		
Model A3		Text Analysis Result & Schema Mode		
Model A4		Text Analysis Result & Schema Mode	Text Analysis Result * Schema Mode	
Model B0	SMI Result	None	None	Participants
Model B1		Final Agent Result		
Model B2		Schema Mode		
Model B3		Final Agent Result & Schema Mode		
Model B4		Final Agent Result & Schema Mode	Final Agent Result * Schema Mode	

TABLE IV: Overview of the models used in the analysis of H1a and H1b. Models used for H1a start with an A, and models for H1b start with a B.

a non-inferiority margin, for which it must hold that two results within that margin are not considered inferior to each other. To obtain this margin, we used the table presented in Sauro et al's book [38]. They provide different grading scales, one of which is the standard American letter grading scale (A-F)[9]. From this, the smallest range of a full letter grade (the B grade ranges from 72.6 to 78.8 points, leading to a difference of 6.2 points on a SUS score) was taken. This 6.2 points margin was then compared to the lower limit of the 95% confidence interval of the difference between the SUS scores for the agent and the SMI questionnaire. A positive difference means the the participant rated the agent higher than the SMI questionnaire, meaning that if the lower limit of the 95% confidence interval is higher than -6.2, the agent can be concluded to be non-inferior to the SMI questionnaire. Figure 5, taken from [39], shows a graphical representation of the non-inferiority concept.

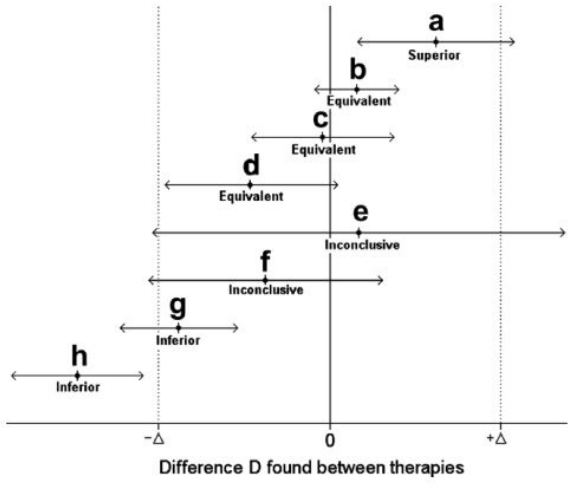

Fig. 5: Hypothetical scenarios of the confidence intervals of observed treatment differences. In this experiment, we evaluate the non-inferiority, which in this image would be a difference like a, b, c or d. The non-inferiority margin Δ in this case is 6.2.

Finally, the time spent on interacting with the conversational agent and filling in the SMI questionnaire was analysed using a two-tailed paired t-test. The p-value used was 0.05.

[9]For more information on how letter grading compares to numeric grading, see https://nces.ed.gov/nationsreportcard/hsts/howgpa.aspx

H. Results

Schema mode	TP	FP	FN	TN
Angry	160	81	4	1
Detached	22	10	17	30
Happy	3	0	44	25
Healthy	126	5	28	2
Impulsive	40	99	12	19
Punishing	13	27	19	27
Vulnerable	0	0	38	27

TABLE V: Results for H1a. A True Positive (TP) indicates that a schema mode was selected by the text analysis algorithm, and that it was consequently confirmed by the follow-up questions. A False Positive (FP) indicates that it was selected by the text analysis algorithm, but that the follow-up questions did not confirm it. A False Negative (FN) indicates that it was randomly selected, and confirmed by the follow-up questions, and a True Negative (TN) indicates that a schema mode was randomly selected, and not confirmed by the follow-up questions.

1) H1a: Table V shows the confusion matrix of the internal performance. We see that the quality of the text analysis algorithm seems to differ per schema mode. For example, for the detached schema mode, being in the text analysis result seems to increase the odds of being in the final agent result (22/10 for the ones in the text analysis result vs 17/30 for the randomly added ones). However, when looking at the angry schema mode, the odds when present in the text analysis result are roughly 2/1 (160/81), but when it is not in the text analysis result, this is higher, at 4/1.

If we then look at Table VI, we see whether there was a significant difference between the models' explaining power. We see that all models significantly improve compared to their relative null-models (i.e. the model without the particular effect mentioned). Table VII shows the coefficients of model A4, and whether they are significant. While the model comparison shows that adding these predictors to a model improves its ability to fit the data, none of the coefficients are significant. This seems counter-intuitive at first, but becomes easier to understand when looking at Figure 6. What we see is that the effect of being included in the result of the text analysis differs per schema mode. For instance, for the angry schema mode,

which is the baseline condition in Table VII, the odds of being in the final agent result actually decrease when it is included in the text analysis result. This matches our observations from Table V.

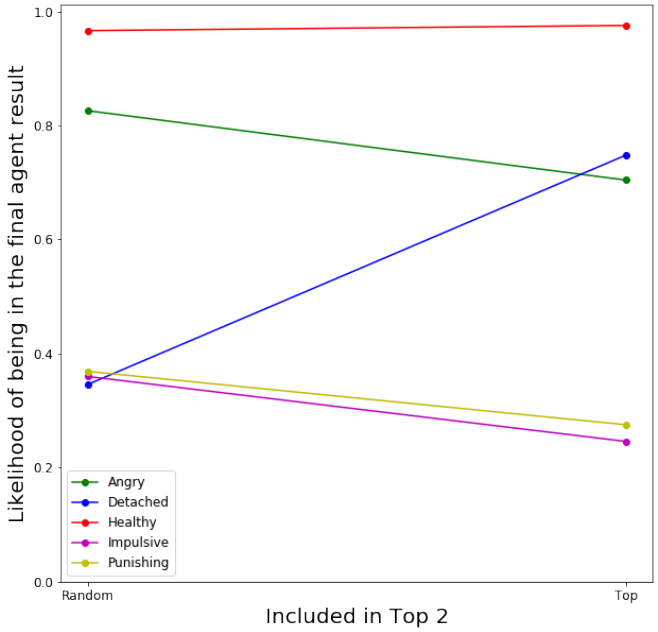

Fig. 6: The plot of the interaction effect of the different schema modes, extracted from Model A4.

Predictor variable (models compared)	χ^2(df)	*P* value
Text analysis (A0 vs A1)	6.64(1)	0.01
Schema mode (A0 vs A2)	208.26(4)	<0.001
Interaction effect (A3 vs A4)	10.42(4)	0.03

TABLE VI: Results of the ANOVA comparison between models including a particular predictor variable and their respective null model.

Parameter	Odds Ratio	Standard Error	z value	*P* value
Intercept	4.74	1.34	1.16	0.24
Text analysis	0.50	1.35	-0.51	0.60
Detached schema mode	0.11	1.39	-1.58	0.11
Healthy schema mode	6.04	1.55	1.16	0.25
Impulsive schema mode	0.11	1.41	-1.51	0.13
Punishing schema mode	0.12	1.39	-1.51	0.13
Text analysis*detached	11.19	1.49	1.62	0.11
Text analysis*healthy	2.75	1.63	0.62	0.53
Text analysis*impulsive	1.16	1.44	0.10	0.92
Text analysis*punishing	1.29	1.47	0.18	0.86

TABLE VII: The fixed effects of different predictor variables in model A4. The intercept is the Angry schema mode.

2) H1b: Table VIII shows the comparisons of the models relating to the external performance. It shows that adding the final agent result and schema mode as predictors to the model both improve the model's ability to fit the data, but that adding an interaction effect does not seem to improve it significantly. When the looking at Table IX, we see the coefficients of model B4, where the intercept is again the angry schema mode. Figure 7 shows the plot of the interaction effect of the different modes.

As we can see in Table IX, the coefficient for the effect of being in the final agent result on the SMI result is 5.20 with a p-value of less than 0.001. This means that the odds of a schema mode in the final agent result to also be in the SMI result, are 5.2 times higher than if it was not in the final agent result. This matches the observations in Figure 7, as the slope of all the lines is positive, indicating that independent of the schema mode, the odds of being included in the SMI result increases when included in the final agent result. The interaction effect for the detached schema mode differs significantly from the baseline angry schema mode, which indicates that the strength of this effect is different for different schema modes. This again matches the expectation from Figure 7, as the slope of the "detached" line is a lot less steep than the slope for the "angry" line. The significant differences of some of the schema modes indicate that the intercepts of the lines are different from the angry schema mode, which again is not a surprise when looking at Figure 7. For example, the intercept of the "healthy" line is higher than the "angry" line, and the "punishing" line is a lot closer to "angry".

Predictor variable (models compared)	χ^2(df)	*P* value
Agent result (B0 vs B1)	175.71(1)	<0.001
Schema mode (B0 vs B2)	644.5(6)	<0.001
Interaction effect (B3 vs B4)	9.21(6)	0.16

TABLE VIII: Results of the ANOVA comparison between models including a particular predictor variable and their respective null model.

Parameter	Odds Ratio	Standard Error	z value	*P* value
Intercept	0.25	0.24	-5.78	<0.001
Agent result	5.20	0.30	5.56	<0.001
Detached schema mode	1.66	0.28	1.84	0.07
Happy schema mode	10.29	0.28	8.34	<0.001
Healthy schema mode	61.20	0.41	10.11	<0.001
Impulsive schema mode	0.53	0.30	-2.09	0.04
Punishing schema mode	1.14	0.28	0.45	0.65
Vulnerable schema mode	1.79	0.27	2.13	0.03
Agent result*detached	0.38	0.48	-2.02	0.04
Agent result*happy	1.38	0.65	0.50	0.62
Agent result*healthy	1.31	0.85	0.32	0.75
Agent result*impulsive	1.29	0.47	0.55	0.58
Agent result*punishing	0.47	0.51	-1.47	0.14
Agent result*vulnerable	0.90	0.50	-0.22	0.83

TABLE IX: The fixed effects of different predictor variables in model B4. The intercept is the Angry schema mode.

3) H2: The mean SUS score of the conversational agent was 73.8, and the mean SUS score for the SMI questionnaire was 79.7. The 95% confidence interval of the difference between the two was (-7.8, -3.9), meaning its lower limit is lower than the -6.2 points limit required for concluding non-inferiority.

4) H3: The analysis of the time that a participant spent interacting with the conversational agent (M = 413.9 seconds, SD = 265.9 seconds) and the time spent filling in the SMI questions (M = 419.4 seconds, SD = 486.3 seconds) indicated

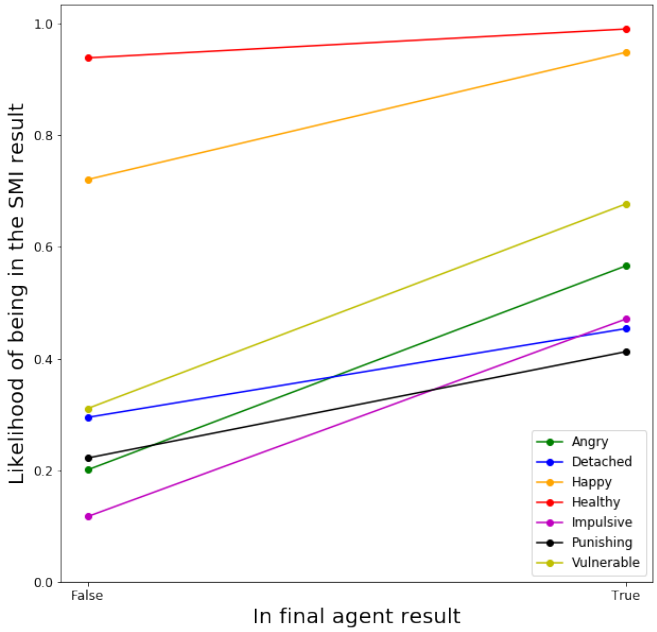

Fig. 7: The plot of the interaction effect of the different schema modes, extracted from model B4.

that there was no significant difference between the two, t(292) = 0.18, p = 0.86.

I. Exploration

Unfortunately, some participants experienced issues during the interaction with the conversational agent. We hypothesized that these issues negatively impacted their usability ratings of the conversational agent, as well as their time spent interacting with the conversational agent. As a result, we marked the conversations where this happened, and ran the analyses of H2 and H3 again, this time controlling for the error.

For the difference in usability (H2), the results of this analysis can be found in Table X. As can be seen, the mean of the scores for the SMI questionnaire are relatively constant, but the participants with issues rated the agent almost 10 points lower than the people without issues. This is reflected in the confidence intervals too, which can be seen in Figure 8. Comparing this to Figure 5, it can be seen that while the total confidence interval does not stay within the -6.2 boundary for non-inferiority, the confidence interval for the participants without issues does, and the confidence interval for participants with errors lies completely outside of that boundary.

	Rows	Mean agent	Mean SMI	95% CI
Total	293	73.8	79.7	(-7.8, -3.9)
No issues	155	78.1	79.6	(-3.7, 0.6)
Error	138	69.1	79.8	(-13.9, -7.5)

TABLE X: The results of explorative analysis of the SUS scores.

For the difference in time (H3), these results can be found in Table XI. While the means of the time spent on either task are close together in the total sample, it can also be seen that

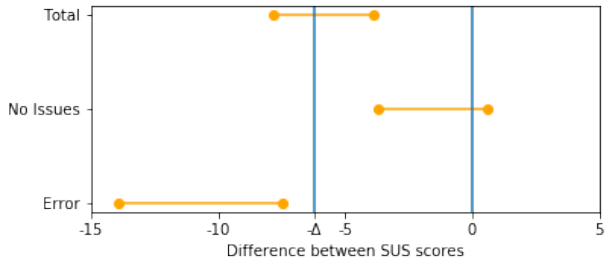

Fig. 8: The confidence intervals of the difference in usability scores. Negative values imply a higher rating for the SMI questionnaire than the conversational agent, and the -Δ is set at -6.2.

on average, people who experienced issues spent almost 200 seconds more on the interaction with the agent than people who did not experience those issues. For the SMI questionnaire it is reversed: people without issues spent almost 100 seconds longer on the questionnaire than people who did experience errors.

	Rows	Mean agent	Mean SMI	T-test T	T-test p-value
Total	293	413.9	419.4	0.18	0.857
No issues	155	319.7	463.4	2.82	0.005
Error	138	519.6	370.1	-5.94	<0.001

TABLE XI: The results of the explorative analysis regarding the time spent on each task, in seconds.

For more in-depth analysis of these issues as well as more details on the data preparation regarding this exploration, please see Appendix C.

J. Discussion and conclusions

In this study, we evaluated a conversational agent designed to assess a person's schema modes. We evaluated the internal and external performance, as well as its usability and time required when compared to the current standard of assessment.

In terms of internal performance, the results show that while adding the text analysis result and the schema mode improve the model's ability to fit the data, the direction of the effect differs per schema mode. This is because of the interaction effect, which is best shown in Figure 6. It is possible that this is partially influenced by the skewedness of the text analysis algorithm discussed in Appendix B. Since some schema modes are almost always highly ranked by the text analysis algorithm and some rarely, there is no independence of errors between the text analysis result and schema mode variables. Because of this, it is not possible to say how knowing the text analysis result changes the odds of being in the final agent result, as whether it increases or decreases the odds depends too much on which schema mode it pertains to.

However, in terms of external performance, the results show that the final agent result is a significant predictor of whether a schema mode will be confirmed by the SMI questionnaire. On average, knowing that a schema mode is confirmed by the conversational agent raises the odds of it being confirmed by the SMI questionnaire by 5.20 times, with a p-value of <0.001.

The strength of this effect varies per schema mode, but is always positive. The schema mode as a predictor varies in strength, and especially the healthy and happy schema modes are likely to be confirmed by the SMI questionnaire. This prevalence matches with literature [10], and likely has to do with the fact that the participant sample was composed of the general public, and as such the majority can be expected to be psychologically stable.

The initial results show that the usability of the agent is rated as significantly lower than that of the SMI questionnaire. On closer inspection however, the results show that when controlling for errors in implementation, the usability of the agent can be accepted as non-inferior to the SMI. It is important to note that these findings may not be generalizable, as participants with no errors were a non-random subsample. The results do hint at that if these errors in implementation were fixed, that the resulting agent might be considered non-inferior, and therefore suggest that an agent-based method has merit.

Regarding the time required for an interaction, the results show that there is no significant difference between the agent and the SMI. Yet again on closer inspection, the results show that when controlling for errors in implementation, there is a significant difference between the time spent on interacting with the conversational agent and filling in the SMI questionnaire, with the conversational agent taking less time. The same caveats as with the usability analysis apply here, and so while it cannot be concluded that the conversational agent is faster than the SMI questionnaire, the results seem to indicate that there is promise to the concept, if these issues were patched.

V. Discussion

A. Conclusion

The primary research question for this research was: *How can a conversational agent be used to assess a person's schema modes?* To answer this question, it was broken down in several sub-questions:

1) What should the design requirements be for a conversational agent to be able to assess a person's schema modes?: A conversational agent intended to assess a person's schema modes should meet several design requirements. Discussion with a panel of experts revealed there is a demand for a new measure different than the SMI questionnaire, specifically a measure that can also analyse open information and adapt to user input. However, this measure should not require continuous monitoring, and instead be able to function independently. Additionally, given that the agent will be maintained by therapists that might lack a technical background, operation and extracting knowledge from the agent should require as little technical knowledge as possible.

2) How should these design requirements be implemented to realize a conversational agent that can assess a person's schema modes?: A conversational agent was implemented to assess a person's schema mode in several stages. In the first stage the person was asked to retell a recent emotional event, from which the agent distilled a list of schema modes related to that story. This list was used to form a selection of questions from the SMI questionnaire, which were asked in the second stage. During this stage, the agent used an adaptive questionnaire method to reduce the amount of questions that needed to be asked. The answers of these questions finally culminated in an assessment of the schema modes.

3) How well can a conversational agent be used to assess a person's schema modes?: Evaluation of the agent showed that the agent was able to predict a person's schema modes. The initial prediction based on the text analysis of the recent emotional story was shown to be a significant predictor for whether a schema mode would be confirmed by the follow-up questions, and a schema mode that was confirmed by the conversational agent was 5.20 times more likely to be confirmed by the SMI questionnaire too. The findings also suggest that the conversational agent does not seem to be considered less user-friendly than the SMI questionnaire, and seems to be faster to use too, though due to implementation errors, more testing is required to confirm this.

B. Limitations

There are some limitations to this work. First of all, the text analysis algorithm is currently lacking. In its current state, it barely predicts 2 of 7 schema modes. While the efficacy of the text analysis algorithm was not the main point of this work, there are many different text analysis algorithms [40], and it is likely that the overall efficacy of the conversational agent would improve with better classification. Additionally, the evaluation of the agent was done on members of the general public, and the hypothesized usability and time saved did not occur for the full sample. People with mental health issues could have significantly different issues compared to the group used for evaluation.

C. Contributions

This work is a new approach to assessing a person's schema modes. Instead of using a rigid and long questionnaire followed by significant human labour afterwards, the conversational agent uses a combination of open and closed questions in an adaptive way, providing therapists with a different measure of a person's schema modes. Another benefit is that while the SMI questionnaire is usually only used once and therefore provides a static assessment, the conversational agent can be interacted with multiple times, providing a therapist more valuable dynamic data in a way that does not require individual therapy time. While not customer-ready, this work serves as a proof of concept that this approach might work when fully developed, and it might even be possible to generalize this approach to be used in different contexts of (mental) health evaluation where the current standard is an interview and/or a questionnaire. The final contribution of this work is that the data gathered in this research is available as a public dataset. This dataset can be used for training conversational agents or otherwise improving knowledge and approaches in the context of schema modes.

D. Future research

There are several areas where this work would benefit from more research. As mentioned in the limitations section, there are several possible improvements in terms of quality of automatic classification. Both the algorithm itself as well as the data used to train the algorithm can be expanded upon. While the algorithm now uses a static top-n approach to select which schema modes will be evaluated by the conversational agent, an approach based on confidence levels or another less rigid approach could be explored, such as the ConveRSE framework [41].

Another area that could be explored is how the data produced in this research can be used to improve the recommendations of the algorithm. While at the start of this work there was no corpus that could be used for training, such a corpus now exists. By using the results of SMI questionnaire to classify the stories written by the participants, the data in this research can be used for iterative improvement of the conversational agent.

In terms of context, there are also several different areas for future research. The expansion to more schema modes is an obvious one, which can be done in conjunction with developing a new algorithm for classification and selection. Another interesting question is how this conversational agent can be further developed to support longitudinal research. This study right now has only focused on a single observation moment, but having multiple moments of evaluation spread out over time will likely generate new insights, which is a step further in the direction of vision for this project. Finally, while this project has attempted to get rid of the artefacts of the pen-and-paper questionnaire approach, one more area of possible future research would be eliminating another artefact, and stepping beyond just text-based evaluation. A computerized system can take in information in more modes than just text, and there is research [42] on emotion recognition based on things like typing speed and pressure on a mobile device. It would be very interesting to see in what way this data could complement the current classification process of the conversational agent.

E. Final remarks

This work has shown a novel way to assess a person's schema modes. It has shown that while there is room for improvement, the concept of using a conversational agent for this assessment has merit. Considering the current events in the world, the mental health crisis is unlikely to slow down, and this technology can provide the therapist community with an additional method to help their patients, while at the same time reducing their own workload.

APPENDICES

APPENDIX A
EVALUATION DESIGN BACKGROUND

In this appendix we will elaborate on the design rationale for the evaluation metrics. How do we evaluate a conversational agent? This was a non-trivial question, and needed specification, for which we first looked at literature.

One recurrent theme when reviewing literature was the lack of large sample sizes. It seemed to be a near-universal recommendation from regular and review papers alike [20], [21], [24]. A major limiting factor of this was the recruitment of participants and running the experiment in person. Because of that, we decided to use an online platform for the experiment. Several such platforms exist, with varying degrees of control on which participants to include in the study, which was another recommendation [24].

Furthermore it is important to realize that while conversational agents are becoming more mainstream, most participants will likely not be used to interacting with a conversational agent. Especially when sampling from a wide variety of subjects who interacting with new piece of technology in an unsupervised manner, it is important to introduce the technology to the participant. Elmasri et al [24] recommended introducing the conversational agent to participants with an example conversation, to show the participants what to expect. However, it is equally important to avoid biasing participants. Denecke [22] noted that when using keywords, it was important to not tell participants beforehand what keywords there are, to avoid that participants (sub)consciously use or not use these words in their story.

Additionally, there was the recommendation to test not only for user acceptance of the new technology, but to also test for the actual efficacy at solving the problem it claims to solve [21], [20], [25]. While this may seem like an obvious thing to do, things like behaviour change or psychological mood are subject to many different biases and effects, making defining efficacy a non-trivial problem, let alone measuring it well.

In our own situation, there were several variables that could be measured to determine efficacy. The naive approach would have been comparing it to the current standard, but this was not without complication. The conversational agent asked questions about a recent emotional event, which was a single isolated moment. The SMI on the other hand is designed to get a more long-term picture of an individual's schema modes. As a result, how well the final result of the conversational agent correlated with the result of the SMI was not only influenced by the quality of the conversational agent, but also by how well the emotional event the person told the conversational agent about represented their personality on a larger scale. It is not hard to imagine that a person who otherwise felt perfectly robust and secure, recently experienced a situation that made them feel vulnerable, and that they chose to tell that story to the conversational agent. In that situation, while the conversational agent might correctly recognize that story as belonging to a vulnerable person, that sentiment may not show up at all in their SMI result.

As such, we defined the performance of the agent in 2 ways: The internal performance of the conversational agent, which was determined by how often the predicted schema modes of the text analysis algorithm of the conversational agent were confirmed by the follow-up questions related to them, and secondly the external performance, which was how well the final result of the conversational agent compared to the result of the SMI.

Furthermore, efficacy in this situation was not just limited

to how accurate the conversational agent was when it came to recognizing schema modes. Research shows that people perform better in terms of accuracy and speed [33] in dynamic systems, and another motivation for this work was to save therapists and patients alike time and effort. Therefore, an evaluation of the conversational agent also needed to reflect this aspect. While effort is harder to measure, time spent was easier. As such, we also measured the amount of time it took participants to complete their interaction with the conversational agent, as well as the time it took them to fill in the SMI.

Finally, while user experience alone must not be mistaken for efficacy, it still is an important aspect of the overall efficacy of a tool. Research shows that people prefer conversational agents over human interviewers when talking about highly sensitive topics that are likely to evoke negative self-admissions [31], and that people prefer interactive over static tasks [32]. However, there was already an established way of determining a person's schema modes, and if users disliked the conversational agent so much that they would not want to use it, any potential value it may have provided would be negated. As a result, users should not dislike the conversational agent significantly more than the current standard way of doing it, filling in the SMI.

The above observations led to the hypotheses listed in Section IV.

APPENDIX B
BEFORE/AFTER ALGORITHM ISSUES

As mentioned at the end of Section III, there was a forced configuration change in the Rasa stack, which occurred between the final verification testing and the experiment of the conversational agent, causing unexpected results. In this appendix, we will describe the cause and probable effects of this configuration change.

On September 28th 2020, the PolyAI team decided to take down the ConveRT models from the public domain[10]. This created an issue for many conversational agents implemented in Rasa[11], as their default recommendation for a pipeline included ConveRTTokenizer and ConveRTFeaturizer[12]. After this became known, a Rasa employee recommended a different pipeline in the Github issue thread.

Before the models were taken down, our conversational agent also used the ConveRT models, and after they were taken down we switched to the pipeline recommended by Rasa. Regrettably, testing after the switch was not thorough enough to realize the impact of this on the text analysis algorithm. Table XII contains the analysis results from before and after the configuration change. Looking at these results, one important thing to note is that when testing with all vignettes, the text analysis algorithm never selected either the happy or vulnerable schema modes to be in the top 3 of the ranking of all 7 schema modes. This shows that in the setup after the configuration change, the vulnerable and happy schema modes had almost no chance to be included in the text analysis result, something which was also observed in the experiment (See Table V).

	Before		After	
Schema mode	Confidence	Top 3	Confidence	Top 3
Angry	0.49	71%	0.14	64%
Detached	0.17	7%	0.53	57%
Happy	0.15	7%	0.08	0%
Healthy	0.12	36%	0.43	36%
Impulsive	0.16	14%	0.86	86%
Punishing	0.32	64%	0.80	57%
Vulnerable	0.28	100%	0.02	0%

TABLE XII: Vignette evaluation results before and after configuration change. The confidence is the average confidence score on a scale from 0 (does not fit at all) to 1 (fits perfectly), assigned by the text analysis algorithm to vignettes that were written specifically to invoke that schema mode. The Top 3 is how often a particular schema mode appeared in the top 3 of the confidence ranking of the text analysis algorithm when looking at all vignettes.

In conclusion, after the configuration change, the average ranking of the text analysis algorithm had less variety, and rarely included 2 out of 7 schema modes in the top 3. Additionally, as can be seen in Section IV, the effect of recommendation differed per schema mode, occasionally increasing the odds of being confirmed by the follow-up questions, while decreasing those odds for other schema modes. From this it becomes evident that the text analysis algorithm requires more research. Not recommending a schema mode at all is problematic, especially considering that in the experiment according to the SMI results, 37% of participants qualified for the vulnerable schema mode, and 74% qualified for the happy schema mode.

APPENDIX C
EXPLORATION

As mentioned in Section IV, some issues were discovered during the analysis process, which led to unexpected results. In this appendix, we will discuss 2 main issues and their effects on the analysis results.

A. First main issue

The first issue was that though the instructions told the participants to quit after the final text was received, some participants did not do so, as can be seen in Figure 9. Due to an uncaught error in implementation, the conversational agent saw this as another cue to analyse a recent emotional story, and started the process of analysing this "recent emotional story" anew, accompanied with questions. While some users who had this issue ignored this error and just continued with the experiment, some answered the follow-up questions, with some repeating this process several times. We hypothesized that this issue did not only negatively impact the time spent by the user, but also decreased the perceived usability of the conversational agent.

[10]https://github.com/PolyAI-LDN/polyai-models

[11]https://github.com/RasaHQ/rasa/issues/6806

[12]https://legacy-docs-v1.rasa.com/1.10.23/nlu/choosing-a-pipeline/Note that at time of reading (01-03-2021) this recommendation is still there in the latest version of the documentation, and has not changed, despite 9 updates and 5 months since the ConveRT models were taken offline.

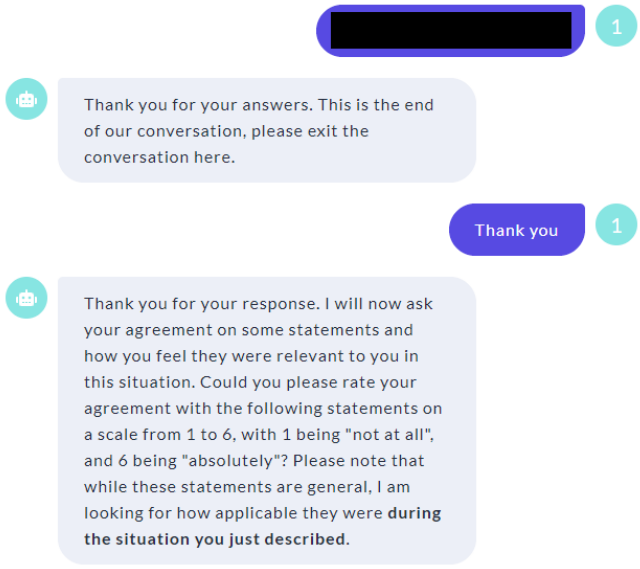

Fig. 9: A user answers the last question of the conversational agent (answer redacted for privacy purposes), after which the conversational agent asks them to exit the conversation. The user instead responds with "Thank you", which the conversational agent misinterprets as the start of another conversation.

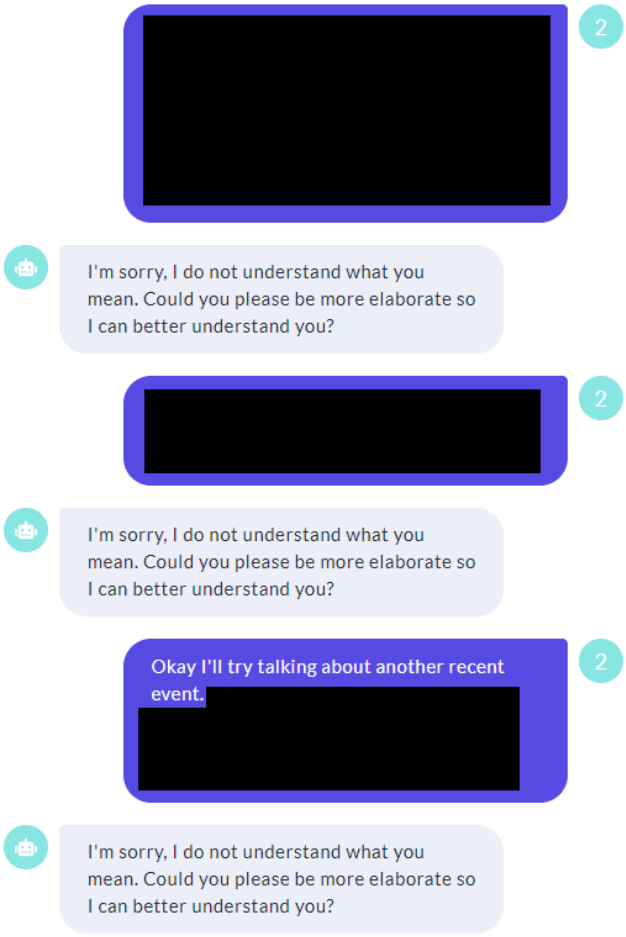

Fig. 10: After already having attempted to write their story once, the conversational agent continues to not understand the user. The user eventually switches to a different recent emotional event, which is also not immediately successful.

B. Second main issue

The second main issue was due to the setup of the automatic classification. In the case where the confidence of the prediction was not high enough after the analysis of the recent emotional story, a fallback action was implemented in the system, where the conversational agent would ask the participant to elaborate on their story with more detail. This was implemented to avoid accidentally accepting nonsense submissions. While this worked acceptably in testing, during the experiment it became clear that the threshold value for the confidence was set far too high. This also had to do with the issues described in Appendix B. As can be seen in Figure 10, some participants had to re-write their recent emotional story again, in some cases more than 5 times. There were cases where after several unsuccessful tries, the participant decided to tell a different story altogether.

Understandably, there were a lot of comments in the post-experiment questionnaire that indicated frustration with this issue. As a result we hypothesized that this issue negatively impacted the perceived usability of the conversational agent. Additionally, we hypothesized that participants who experienced this issue will have spent more time on the conversational agent due to the time required to re-write their recent emotional story multiple times.

C. Data preparation

After the manual inspection of the conversations, certain hallmarks present in the raw conversation data allowed us to automatically mark the conversations where these issues occurred. This yielded that out of the 293 submissions, 155 participants experienced no issues, 19 participants experienced only the first issue, 103 participants experienced only the second issue, and 16 participants experienced both issues. For clarity, the first issue of talking to the conversational agent after the conversation was supposed to end will be referred to as "ending issue", and the second issue where participants had to rewrite their story multiple times due to the threshold being too high will be referred to as "threshold issue".

D. Effects on H2

For the perceived usability in terms of SUS scores, the histograms shown in Figure 11 were drawn. While the distributions are similar, it can be seen that especially for the participants experiencing the threshold issue the difference in scores skews more negative. The participants with no issues skew quite positive. The subgroups were then analysed per group, from which the results can be found in Table XIII. As we can see, while for the whole group the lower limit of the confidence interval was lower than -6.2, if we look at the group of participants who did not experience implementation errors of the conversational agent this is not the case.

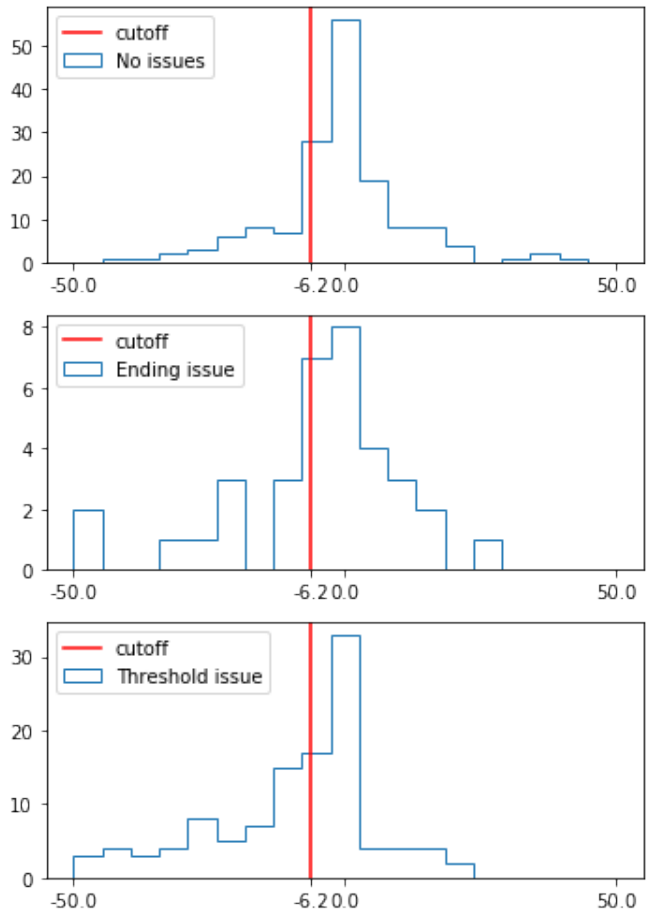

Fig. 11: The histograms of the difference of the SUS scores. Positive scores indicate a higher score for the conversational agent than for the SMI. The red line indicates the -6.2 cutoff point indicated in the hypotheses.

	Rows	Mean agent	Mean SMI	95% CI
No issues	155	78.1	79.6	(-3.7, 0.6)
Ending	35	72.4	77.2	(-10.3, 0.6)
Threshold	119	67.3	79.5	(-15.6, -8.6)
Total	293	73.8	79.7	(-7.8, -3.9)

TABLE XIII: The results of further analysis regarding H2.

Additionally, a general linear model was trained with whether there was an ending or threshold issue as a predictor variable for the difference in SUS scores, in an attempt to isolate the influence of experiencing these issues. From this, 95% confidence intervals were also calculated. The results of this can be found in Table XIV. As can be seen, the estimates for the coefficients show a clear difference, with the threshold issue group being significantly different from the group with no issues, which is also reflected in the confidence intervals. When comparing these with Table XIII the numbers for the ending issue might be surprising, but note that out of the 35 participants who experienced an ending issue, 16 also experienced a threshold issue. When isolated, having an ending issue does not seem to significantly affect the difference in SUS scoring.

	Estimate	Std. Error	t value	p value	95% CI
No issues	-1.7	1.3	-1.4	0.17	(-4.2, 0.7)
Ending	1.7	2.9	0.6	0.55	(-4.0, 7.5)
Threshold	-10.6	1.9	-5.5	<0.001	(-14.4, -6.8)

TABLE XIV: The results of the model trained on the data regarding H2 and including the group of what issue they experienced.

E. Effects on H3

In regards to the time spent by the participants on the conversational agent and the SMI, things are a little bit more complicated. While the SUS scale is a scale from 0-100, the time spent by participants on either the conversational agent or the SMI did not have an upper bound. This lead to some significant outliers, as can be seen in the boxplot of the total experiment time in Figure 12.

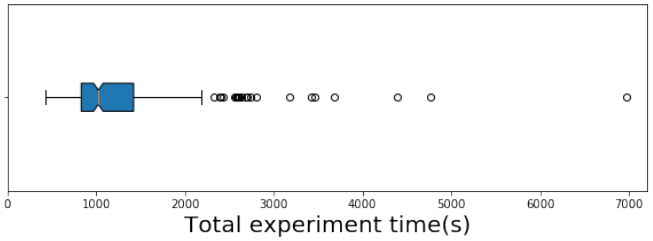

Fig. 12: Distribution of total experiment time. Note that there are some outliers which required almost 2 hours.

When inspecting the means of the different subgroups, the difference becomes quite clear. As can be seen in Table XV, while the mean agent time and mean SMI time are not significantly different when looking at the total, the mean SMI time is significantly higher than the mean agent time for both of the groups that experienced issues. In the group with no issues, this is reversed.

	Rows	Mean agent	Mean SMI	T-test T	T-test p-value
No issues	155	319.7	463.4	2.82	0.005
Ending	35	522.6	370.3	-2.51	0.017
Threshold	119	519.8	366.7	-3.16	<0.001
Total	293	413.9	419.4	0.18	0.86

TABLE XV: The results of further analysis regarding H3, grouped by what issue they experienced.

Again a general linear model was trained with having an ending or threshold issue as a predictor variable for the difference in time. Positive values indicate more time spent on the SMI than on the conversational agent. From this, 95% confidence intervals were also calculated. These can be found in Table XVI.

	Estimate	Std. Error	t value	p value	95% CI
No issues	-132	40	-3.3	<0.001	(-210, -54)
Ending	164	91	1.8	0.073	(-15, 343)
Threshold	263	60	4.4	<0.001	(145, 381)

TABLE XVI: The results of the model trained on the data regarding H3 and including the group of what issue they experienced.

The estimates show that people with no issues spend on average more than 2 minutes less on the conversational agent than on the SMI, while people who experienced threshold issues spent more than 4 minutes longer on the conversational agent. Both differences are significant with $p<0.05$. While the estimate for people with ending issues is that they spend 2.5 minutes longer on the conversational agent, this result is not significant, likely due to the smaller relative sample size. The confidence intervals also reflect these observations.

F. Discussion

As can be seen in the data, the issues mentioned seem to have negatively impacted the time spent and usability scores of the participants that experienced them. Particularly the prevalent threshold issue seems to have impacted participants' usability scores and time spent on the agent, with even the upper limit of the 95% confidence interval being lower than -6.2, and taking more than 4 minutes longer interacting with the agent, versus 2 minutes fewer than when experiencing no issues.

An important limitation of this post-analysis however is that the method of sampling is influenced by the variable that is being measured. If one considers the participants in 2 groups, divided on their skill of interacting with a conversational agent, either through experience or inherent aptitude. For this experiment, it is likely that the group with lower skill will run into the errors more often that the participants with more skill. This is especially true for the ending issue, where participants who followed the instruction of the conversational agent and exited the conversation when told to do so did not experience the issue. It is therefore not possible to conclude that these errors alone are responsible for the difference in scores, and that if they were fixed, the same experiment would produce sample scores seen here for the group that experienced no issues. Further testing would be required if patching these issues would bring the sample scores to the level where the hypotheses could be considered confirmed.

APPENDIX D
RASA CONFIGURATION FILE

```
   language: en
pipeline:
  - name: WhitespaceTokenizer
  - name: RegexFeaturizer
  - name: LexicalSyntacticFeaturizer
  - name: CountVectorsFeaturizer
  - name: CountVectorsFeaturizer
    analyzer: "char_wb"
    min_ngram: 1
    max_ngram: 4
  - name: DIETClassifier
    epochs: 100
  - name: EntitySynonymMapper
  - name: ResponseSelector
    epochs: 100
policies:
- name: MemoizationPolicy
- name: TEDPolicy
- name: MappingPolicy
- name: FormPolicy
- name: FallbackPolicy
  nlu_threshold: 0.4
  core_threshold: 0.3
- name: AugmentedMemoizationPolicy
  max_history: 0
  priority: 2.5
```

APPENDIX E
RASA INTENT LIST

```
## intent:happy_child
- I feel loved
- I feel accepted
- I am accepted
- I am satisfied
- I feel calm
- I feel connected to other people
- Belonging
- I have stability in my life
- I have certainty in my life
- I trust people
- I feel safe
- I feel heard
- I am understood
- I am supported
- I am optimistic
- I am spontaneous

## intent:vulnerable_child
- I feel worthless
- I feel inadequate
- I am not enough
- I am lost
- I feel lost
- I am desparate
- Desparation
- I am lonely
- loneliness
- I feel humiliated
- humiliation
- I feel weak
- I am helpless
- I am alone
- I feel left out
- Nobody loves me
- Nobody likes me
- I am inadequate
- I am broken
- I am excluded
- I feel powerless
- It is never enough
- I am not good
- I am a mess
- Pathetic
- My future is bleak
- I have no future
- Rejection
- I need help
- I am ashamed of myself
- I'm scared
- Fear
- I am needy
- Needy
- I am overwhelmed
- I am nervous

## intent:angry_child
- I have to fight
- I am angry
```

```
- I am furious at someone
- I hold on to my anger
- I am furious
- People are with me or against me
- I am angry at someone because they left me
- It makes me angry when someone tells me what
    to do
- I want to punish people for how they treated
    me
- I feel cheated
- I feel treated unfairly
- I want to hurt someone for what they did to
   me
- I want to fight
- People are trying to limit me
- I have a lot of anger inside me
- I have to let my anger go

## intent:impulsive_child
- I have trouble controlling myself
- I act first and think later
- I cannot control my impulses
- I follow my feelings
- I follow my emotions
- I get in trouble because of impulsiveness
- I do not think of consequences
- I say what I feel
- I do things impulsively
- I do first and act later
- I dont think about my actions
- I hurt people by not thinking about what I
   do
- I do not think
- I regret breaking rules
- I just do
- Without thinking
- I did not think

## intent:detached_protector
- I feel flat
- I do not feel anything
- I do not feel connected
- I do not feel my emotions
- I feel nothing
- I dont care about anything
- Nothing matters to me
- I feel distant from other people
- I feel cold
- I feel emotionless
- I do not feel connected to other people
- I do not feel connected to myself
- I am indifferent
- I dont want to feel
- I don't like to feel
- I don't want to
- It is not necessary
- I don't think it helps
- It doesn't matter
- I don't need it

## intent:punishing_parent
- I do not deserve fun
- I do not deserve enjoyment
- I do not deserve pleasure
- I do not deserve a break
- I punish myself
- Selfharm
- I injure myself
- I am a terrible person
- I am a bad friend
- I am not a good child
- I am an awful parent
- I do not forgive myself
- I am angry at myself
- I dont deserve sympathy
- I do not deserve pity
- I deserve to be punished
- It is my fault
- Bad things are my fault
- I am the cause of my problems
- I am bad
- It is my fault
- I am unsuccessful
- I can't do it anyway
- I should be able to do this
- There is no point
- Disappointing
- I am a disappointment
- I am useless

## intent:healthy_adult
- I can solve my own problems
- I know how to express my emotions
- I can learn
- I can grow
- I can change
- I can stand up for myself
- I can assert what I need
- I know who I am
- I know what I need to be happy
- I can make myself happy
- I am a good person
- I can take care of myself
- I can handle my emotions
- I can handle bad situations
- I can do boring things
- I am happy with myself
- I am proud of myself
- My emotions do not overwhelm me
- I am stable
- I am worth the effort
- I am worth attention
```

REFERENCES

[1] E. Jané-Llopis, P. Anderson, S. Stewart-Brown, K. Weare, K. Wahlbeck, D. McDaid, C. Cooper, and P. Litchfield, "Reducing the silent burden of impaired mental health," *Journal of health communication*, vol. 16, no. sup2, pp. 59–74, 2011.

[2] R. C. Kessler, P. Berglund, O. Demler, R. Jin, K. R. Merikangas, and E. E. Walters, "Lifetime prevalence and age-of-onset distributions of dsm-iv disorders in the national comorbidity survey replication," *Archives of general psychiatry*, vol. 62, no. 6, pp. 593–602, 2005.

[3] S. Torgersen, E. Kringlen, and V. Cramer, "The prevalence of personality disorders in a community sample," *Archives of general psychiatry*, vol. 58, no. 6, pp. 590–596, 2001.

[4] B. F. Grant, D. S. Hasin, F. S. Stinson, D. A. Dawson, S. P. Chou, W. J. Ruan, and R. P. Pickering, "Prevalence, correlates, and disability of personality disorders in the united states: results from the national epidemiologic survey on alcohol and related conditions.," *The Journal of clinical psychiatry*, vol. 65, no. 7, pp. 948–958, 2004.

[5] D. I. Soeteman, R. Verheul, and J. J. Busschbach, "The burden of disease in personality disorders: diagnosis-specific quality of life," *Journal of personality disorders*, vol. 22, no. 3, pp. 259–268, 2008.

[6] D. I. Soeteman, L. H.-v. Roijen, R. Verheul, and J. J. Busschbach, "The economic burden of personality disorders in mental health care.," *Journal of Clinical Psychiatry*, vol. 69, no. 2, p. 259, 2008.

[7] S. McMain and A. E. Pos, "Advances in psychotherapy of personality disorders: a research update," *Current Psychiatry Reports*, vol. 9, no. 1, pp. 46–52, 2007.

[8] J. E. Young, J. S. Klosko, and M. E. Weishaar, *Schema therapy: A practitioner's guide*. Guilford Press, 2003.

[9] J. L. Lebow, *Twenty-first century psychotherapies: Contemporary approaches to theory and practice*. John Wiley & Sons, 2012.

[10] J. Lobbestael, M. van Vreeswijk, P. Spinhoven, E. Schouten, and A. Arntz, "Reliability and validity of the short schema mode inventory (smi)," *Behavioural and Cognitive Psychotherapy*, vol. 38, no. 4, pp. 437–458, 2010.

[11] A. Arntz and G. Jacob, *Schema therapy in practice: An introductory guide to the schema mode approach*. John Wiley & Sons, 2017.

[12] L. L. Bamelis, S. M. Evers, P. Spinhoven, and A. Arntz, "Results of a multicenter randomized controlled trial of the clinical effectiveness of schema therapy for personality disorders," *American Journal of Psychiatry*, vol. 171, no. 3, pp. 305–322, 2014.

[13] A. D. van Asselt, C. D. Dirksen, A. Arntz, J. H. Giesen-Bloo, R. van Dyck, P. Spinhoven, W. van Tilburg, I. P. Kreimers, M. Nadort, and J. L. Severens, "Out-patient psychotherapy for borderline personality disorder: cost-effectiveness of schema-focused therapy v. transference-focused psychotherapy," *The British Journal of Psychiatry*, vol. 192, no. 6, pp. 450–457, 2008.

[14] J. Lake and M. S. Turner, "Urgent need for improved mental health care and a more collaborative model of care," *The Permanente Journal*, vol. 21, 2017.

[15] A. Arntz, J. Klokman, and S. Sieswerda, "An experimental test of the schema mode model of borderline personality disorder," *Journal of behavior therapy and experimental psychiatry*, vol. 36, no. 3, pp. 226–239, 2005.

[16] J. Lobbestael, A. Arntz, M. Cima, and F. Chakhssi, "Effects of induced anger in patients with antisocial personality disorder," *Psychological medicine*, vol. 39, no. 4, p. 557, 2009.

[17] J. Lobbestael and A. Arntz, "Emotional, cognitive and physiological correlates of abuse-related stress in borderline and antisocial personality disorder," *Behaviour research and therapy*, vol. 48, no. 2, pp. 116–124, 2010.

[18] J. Weizenbaum *et al.*, "Eliza—a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[19] B. A. Shawar and E. Atwell, "Different measurements metrics to evaluate a chatbot system," in *Proceedings of the workshop on bridging the gap: Academic and industrial research in dialog technologies*, pp. 89–96, Association for Computational Linguistics, 2007.

[20] D. C. Mohr, M. N. Burns, S. M. Schueller, G. Clarke, and M. Klinkman, "Behavioral intervention technologies: evidence review and recommendations for future research in mental health," *General hospital psychiatry*, vol. 35, no. 4, pp. 332–338, 2013.

[21] D. Bakker, N. Kazantzis, D. Rickwood, and N. Rickard, "Mental health smartphone apps: review and evidence-based recommendations for future developments," *JMIR mental health*, vol. 3, no. 1, p. e7, 2016.

[22] K. Denecke, S. Lutz Hochreutener, A. Pöpel, and R. May, "Talking to ana: A mobile self-anamnesis application with conversational user interface," in *Proceedings of the 2018 International Conference on Digital Health*, pp. 85–89, ACM, 2018.

[23] S. Mujeeb, M. H. Javed, and T. Arshad, "Aquabot: A diagnostic chatbot for achluophobia and autism," *International Journal Of Advanced Computer Science And Applications*, vol. 8, no. 9, pp. 39–46, 2017.

[24] D. Elmasri and A. Maeder, "A conversational agent for an online mental health intervention," in *International Conference on Brain Informatics*, pp. 243–251, Springer, 2016.

[25] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial," *JMIR mental health*, vol. 4, no. 2, p. e19, 2017.

[26] S. Jaiswal, M. Valstar, K. Kusumam, and C. Greenhalgh, "Virtual human questionnaire for analysis of depression, anxiety and personality," in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pp. 81–87, 2019.

[27] A. Ho, J. Hancock, and A. S. Miner, "Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot," *Journal of Communication*, vol. 68, no. 4, pp. 712–733, 2018.

[28] J. D. Robinson, "An interactional structure of medical activities during acute visits and its implications for patients' participation," *Health Communication*, vol. 15, no. 1, pp. 27–59, 2003.

[29] H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, and R. J. Mislevy, *Computerized adaptive testing: A primer*. Routledge, 2000.

[30] M. Van Vreeswijk, J. Broersen, and M. Nadort, *The Wiley-Blackwell handbook of schema therapy: Theory, research, and practice*. John Wiley & Sons, 2012.

[31] M. D. Pickard, C. A. Roster, and Y. Chen, "Revealing sensitive information in personal interviews: Is self-disclosure easier with humans or avatars and under what conditions?," *Computers in Human Behavior*, vol. 65, pp. 23–30, 2016.

[32] A. Locoro, F. Cabitza, R. Actis-Grosso, and C. Batini, "Static and interactive infographics in daily tasks: A value-in-use and quality of interaction user study," *Computers in Human Behavior*, vol. 71, pp. 240–257, 2017.

[33] L. Herman, V. Juřík, Z. Stachoň, D. Vrbík, J. Russnák, and T. Řezník, "Evaluation of user performance in interactive and static 3d maps," *ISPRS International Journal of Geo-Information*, vol. 7, no. 11, p. 415, 2018.

[34] D. Allaart and W.-P. Brinkman, "Assessment of schema modes through a conversational agent," Oct 2020. Available at https://osf.io/kwhcu.

[35] J. Brooke, "Sus: a quick and dirty usability scale," *Usability evaluation in industry*, p. 189, 1996.

[36] A. Bangor, P. Kortum, and J. Miller, "Determining what individual sus scores mean: Adding an adjective rating scale," *Journal of usability studies*, vol. 4, no. 3, pp. 114–123, 2009.

[37] J.-p. Liu, H.-m. Hsueh, E. Hsieh, and J. J. Chen, "Tests for equivalence or non-inferiority for paired binary data," *Statistics in medicine*, vol. 21, no. 2, pp. 231–245, 2002.

[38] J. Sauro and J. R. Lewis, *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann, 2016.

[39] E. Christensen, "Methodology of superiority vs. equivalence trials and non-inferiority trials," *Journal of hepatology*, vol. 46, no. 5, pp. 947–954, 2007.

[40] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams engineering journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

[41] A. Iovine, F. Narducci, and G. Semeraro, "Conversational recommender systems and natural language: A study through the converse framework," *Decision Support Systems*, vol. 131, p. 113250, 2020.

[42] H. Lee, Y. S. Choi, S. Lee, and I. Park, "Towards unobtrusive emotion recognition for affective social communication," in *2012 IEEE Consumer Communications and Networking Conference (CCNC)*, pp. 260–264, IEEE, 2012.