The noticeability of behavioral changes of a conversational agent

An evaluation of an agent-based social skills training system

Ashma Zaheer[1]

Supervisor(s): Willem-Paul Brinkman[1]**, Mohammed Al Owayyed**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Ashma Zaheer
Final project course: CSE3000 Research Project
Thesis committee: Willem-Paul Brinkman, Mohammed Al Owayyed, Elmar Eisemann

Abstract

Helpline counselors can be trained faster using agent-based social skills training systems. These systems utilize conversational agents that simulate interactions with users and provide feedback. This research evaluates the noticeability of behavior changes in such a conversational agent. By examining participants' ability to detect and interpret these changes, valuable insights can be gained to enhance the agent's representation of real-life scenarios and improve its effectiveness as a training tool.

A mixed-methods approach was employed, combining quantitative data using descriptive statistics and qualitative data through content analysis. The quantitative analysis revealed a precision score of 93% and a recall score of 36%, indicating a higher accuracy in correctly identifying behavior changes compared to capturing all instances of change. The qualitative analysis revealed that participants' perceptions fell within four distinct categories: communication style, positive emotion, negative emotion, and tone & attitude. Participants demonstrated heightened attentiveness to changes in positive behavior, particularly instances of happiness expressed by the agent, and displayed less attention towards changes in negative behavior.

In conclusion, this research demonstrates that behavior changes in the conversational agent are noticeable, albeit with varying degrees of attention across different types of behavior. Further exploration with a larger sample size is warranted to ascertain the extent of noticeability. The diverse range of behavior changes indicates the agent's capability to adapt and exhibit visible shifts in behavior. These insights can inform the refinement and development of agent-based training systems, ultimately enhancing the training experiences of chat-based helpline counselors.

1 Introduction

Training communication skills are necessary for many human endeavors, such as patient-doctor communication, training helpline workers on suicide prevention lines, or online bullying support centers. De Kindertelefoon [1] is a Dutch helpline with over 600 volunteers. It provides anonymous and confidential support to children and young people who want to talk about their problems or need someone to listen to them. This anonymity is accomplished by offering a safe space where children and teenagers can freely express themselves without fear of judgment or repercussion. In 2019 alone, they obtained 1,5 million new users between the ages of 12 to 18 years old and reported that 72% of their total talk time was over chat [1].

[1] https://www.kindertelefoon.nl/

Currently, volunteers who want to work at such services require training to acquire the minimum competency and prepare them for the high-stress situations they may face daily over chat. Simulations can be used instead to help meet this growing demand and avoid the additional cost, resources, and time required for training these volunteers [3]. These provide the user with an immersive and engaging situation to gain experience [6]. The systems typically involve the interaction between an actual human learner and an interactive agent or virtual patient (VP) representing another human. In this way, the volunteers can get more acquainted with different scenarios and how to overcome them without influencing their encounters with patients or victims. Using such realistic scenarios in training also helps different volunteers to transfer their knowledge to each other [8].

This research focuses on the scenario of a virtual child contacting the children's helpline to discuss being bullied at school. The purpose is to provide volunteers with training in handling similar situations without exposing them to actual conversations with children experiencing distressing circumstances. Previous work has been done in this area, including the design of a conversational agent called Lilobot based on the Belief-Desire-Intention (BDI) model [2]. However, the quality of this agent, particularly its behavior, needs to be determined.

The main research question of this study is, **'Are the changes in the virtual child's behavior noticeable to the participants?'** Answering this question enhances the agent's representation of an actual child in distress, thereby improving the training effectiveness of volunteers by simulating real-life scenarios. This overarching question is further explored through two sub-questions: *'How noticeable are the changes?'* and *'What specific changes are noticed?'* These sub-questions enable a more detailed examination of the leading research question by breaking it into smaller, more focused components. They aim to understand the extent and nature of the behavioral changes observed by participants during their interactions with the virtual child.

The rest of the paper is organized as follows: section 2 provides a detailed description of the method, including information about the participants, materials used, measurement techniques, and a step-by-step account of the experimental procedure. The data analysis process is also outlined in this section. In section 3, the results of the study are presented. Section 4 offers a reflective analysis of the findings and also highlights the limitations encountered during the research. Section 5 explores the responsible research practices employed, including ethical considerations. The paper concludes with a discussion on future directions and a comprehensive summary of the study's main findings in conclusion section 6.

2 Method

The evaluation method employed in this study utilized a mixed methods research approach, incorporating both qualitative and quantitative research techniques. This approach was chosen to comprehensively address the main research question while tackling the two sub-questions individually with different methods. In this section, the participants' demographic is first examined, followed by the materials used, measured variables, procedure of the experiment, and data analysis of the study.

2.1 Participants

Ten participants were recruited for the study, primarily through acquaintances and friends close to Delft University of Technology. The age range of the participants was between 18 and 24 years. Out of the total participants, there was one female and nine males. Regarding their experience with chatbots, three participants reported frequent usage (more than ten times a month), four said occasional use (2-10 times a month), two reported rare usage (once a month or less), and one participant stated that they had never used a chatbot. Notably, one participant who was not fluent in Dutch was included to assess the viability of including participants with limited Dutch language proficiency. However, it was observed that this participant needed help comprehending Lilobot's responses to the same extent as the fluent Dutch-speaking participants, so their results were discarded. Participants without a solid grasp of the Dutch language, at least in written form, were subsequently excluded from the study. The initial goal was to recruit 20 participants. However, as detailed in section 4.2, certain limitations prevented the achievement of this target sample size.

2.2 Materials

To evaluate the noticeability of the conversational agent's behavior changes, participants must engage in several interactions with the bot, allowing them to observe and comprehend its responses. Furthermore, to establish a common baseline and ensure participants understand how conversations conventionally progress in real life, it is crucial to provide them with some guidance for which the Five Phase Model [10] was incorporated.

LiloBot

To assess behavior changes, participants interacted with a chatbot named Lilobot. The development of Lilobot was undertaken by Sharon Afua Grundmann as part of her thesis [6]. The bot's architecture comprises Rasa, Spring, PostgreSQL, and Azure Blob Storage. Participants interacted with Lilobot through a web page interface, which provided instructions in Dutch on how to engage with the bot. The web page featured a chat widget enabling communication.

The agent's behavior is influenced by the BDI model, one of the cognitive models utilized for designing virtual agents [4]. This framework considers the system as a rational agent that possesses specific mental attitudes representing its informational, motivational, and deliberative states [9]. These mental attitudes collectively play a significant role in shaping the behavior exhibited by the system. Additionally, per the principles of the BDI model, an agent holds beliefs concerning the present state of the world and possesses desires regarding how it would like the state of the world to be. Drawing from these beliefs and desires, the agent then identifies a specific goal as an intention to be pursued and accomplished. By considering the connection between the two previously discussed aspects, it is possible to explore how the agents' beliefs, which serve as the initial step in the BDI model, contribute to the behavior exhibited by the bot. Monitoring the evolution of these beliefs throughout a conversation thus allows us to assess if there was a change in behavior at a certain point in the conversation. To expedite participants' progress through the experiment and accommodate their non-expert status, two beliefs were adjusted by increasing their initial values by 0.1. The purpose of this adjustment was to facilitate their advancement through multiple phases within a reasonable timeframe. Specifically, the beliefs that were altered were B5 ("I think KT understands me") and B6 ("I think KT is interested in my story"). This modification enabled participants to progress faster and engage more effectively with the system.

Guidance regarding conversation

To guide participants on how a conversation should flow between a counselor and a person seeking help, they were provided with the Five Phase Model [10], which serves as a structured framework to familiarize participants with the expected flow of conversations, thus enabling a more informed assessment of the bot's behavior. Each stage of the model was described, highlighting its intended purposes and typical conversational dynamics. This allowed participants to gain a basic understanding of the expected progression in a counseling scenario. The model was provided in a handout containing conversation snippets between a counselor and a person seeking help, along with example phrases for each phase (appendix: A). This resource was available to them throughout the experiment to enhance their understanding.

It is important to note that the topic of these conversations was nail-biting, rather than bullying, aiming to avoid participants using the exact phrases or responses from the provided conversations in the current experimental scenario. Utilizing a different topic ensured that the experiment remained unbiased, and participants focused on evaluating the conversational agent's behavior changes based on their understanding and first-hand experience rather than relying on specific phrases or scripts from the example conversations..

2.3 Measures

In this study, we employed a combination of quantitative and qualitative measures to comprehensively assess the changes in behavior exhibited by the conversational agent. These measures included a questionnaire survey and a belief report, which provided valuable insights into the numerical frequency of behavior changes and the specific types of observed changes.

Questionnaire Survey

Qualitative research methods were employed to measure the specific types of behavior changes noticed. The measurement was done through an open-ended survey so that detailed descriptions of the observed behavior changes could be provided. Each survey was custom to the conversation between the participant and the bot. The approach of open-ended questions allowed participants to offer rich and nuanced insights into their experiences and perceptions. The gathered survey responses were then subjected to content analysis, a systematic method for examining recorded communication, to identify common themes and patterns within the data.

A QuestionnaireService was developed in Java to aid the qualitative measurement. This generated a copy of the transcript in Word, capturing the entire interaction between the participant and the bot. For each pair of prompts within the transcript, a question was printed to create the survey: **'Do you notice a change in behavior here? If so, can you explain what is different?'**. The question was the same throughout the survey. The second half of the question asks the participant to think about the change they noticed in the bot's behavior. Some extra white space was left where the participant's answer was expected. An example snippet of this questionnaire survey can be found in Figure 1, which has been translated from Dutch.

KT: So how do you feel when they call you ugly?
Lilo: I sometimes go to school with a stomachache. I just have to cry just thinking about it. When I get home, I am always very sad.

Question: Do you notice a change in behavior here? If so, can you explain what is different?

Yes. It seems like it may suddenly be getting more comfortable with talking to me and is starting to share much more with each message.

Figure 1: A sample from the questionnaire translated from Dutch, with the transcript dialogues (in black), the question (in orange), and the answer of a participant (in blue). The original page can be found in Appendix B.

Belief Report

Quantitative research was to be conducted to evaluate the changes in the agent's behavior regarding numbers. The frequency with which participants observed changes in the agent's behavior during conversations had to be measured from their answers and compared to the actual number of instances the belief changed. They must be tracked and stored to find the exact number of changes in the beliefs during the conversation. This measurement additionally aided in identifying any discrepancies, such as false positives, where participants perceived changes even though there was no corresponding change in the agent's beliefs.

A ReportBeliefService was developed in Java to facilitate this monitoring of beliefs. Like the QuestionnaireService, this service generates a copy of the conversation in a separate Word document. Within each interaction between LiloBot and the participant, any change in the bot's belief is visually represented, as shown in Figure 2. An upward or downward arrow indicates whether the belief increased or decreased after the specific interaction. The code name (e.g., B4) and the full name of the corresponding belief are displayed, along with the old and new values of the belief, providing insight into the magnitude of the change.

It is worth noting that only 9 out of the bot's 17 beliefs were included in this study. The remaining eight beliefs were excluded as they do not exhibit gradual changes but rely on specific thresholds, triggering a sudden change in value to a

KT: So how do you feel when they call you ugly?
Lilo: I sometimes go to school with a stomachache. I just have to cry just thinking about it. When I get home, I am always very sad.

Belief: ↑ (B6) I think KT is interested in my story. From 0.1 to 0.2

Figure 2: Extract from a belief report translated from Dutch, showing a pair of dialogues (in black) and the belief code, full name , and change in the value of the affected belief (in red.)

minimum or maximum. The beliefs are still accounted for as they have been programmed to increase in groups, e.g. when B4 increases beyond a certain threshold, B16 is set to its maximum value. Consequently, these beliefs were not included in the experimental design, focusing on beliefs that undergo gradual transformations during conversations. This constraint should be considered when interpreting the results and generalizing them to the overall behavior of the conversational agent.

2.4 Procedure

The experimental procedure encompassed three primary stages: preparation, data collection, and additional studies. An overview of the procedure can be seen in Figure 3

Preparation

Before data collection, participants were required to fill in a consent form and a demographic survey encompassing three questions about their age range, gender, and prior experience with chatbots. Subsequently, participants received guidance on how the conversation should flow by referring to the Five Phase Model and the provided example conversation snippets (appendix: A).

During the interaction phase, participants engaged with the conversational bot for 5 to 10 minutes or upon reaching phase 3 of the model. It is important to note that participants were not required to point out any specific observations during the interaction explicitly; instead, they were asked to provide feedback afterward. The conversations could unfold in two different ways: either the bot would leave the conversation dissatisfied, or the participant would be able to generate a satisfactory solution, such as guiding the participant to speak with their teacher regarding the bullying issue. The latter outcome was the desired goal. It is also worth mentioning that participants were not expected to possess expertise in counseling, and their successful completion of the conversation was optional. Instead, participants were encouraged at least to reach phase three of the Five Phase Model. This criterion was deemed sufficient for achieving the research objectives, as it allowed for the generation of adequate transcript lengths where points of behavioral change could be identified and analyzed effectively. Participants were allowed to ask questions, while subtle hints were also provided to guide them on the following steps. Often the participants needed to be made aware of the specific phase within the Five Phase Model they were currently engaged in or whether they were prepared to advance to the subsequent phase, for which they were requested to move on. Furthermore, participants' inputs were monitored to address any punctuation and spelling errors, as the bot's responses were known to be less effective in handling such mistakes.

Data Collection
The data collection stage constituted the most time-intensive segment of the experiment. Following the conversation's conclusion, participants were granted a brief break period, during which a questionnaire survey was generated. Clear instructions were provided to guide participants in responding to the questionnaire, emphasizing the importance of providing detailed explanations regarding the observed changes rather than simply stating whether the bot appeared happy or sad. Furthermore, they were allowed to skip points in the conversation where they did not perceive any noticeable changes. However, whenever they did notice a behavior change, they were instructed to mark 'yes' in the transcript and briefly explain the specific aspect of the change they observed, an example of which can be seen in Figure 1. This allowed participants to identify and highlight variations in the agent's tone, level of trust, or any other notable behavioral differences they discerned during the interaction.

Other Studies
This study was performed in collaboration with other researchers. One researcher dedicated their efforts to assessing the perceived believability of the conversational agent, examining how effectively it simulated a human counselor and whether participants found its behavior convincing or authentic. Additionally, the usability of the feedback provided by the agent at the end of the conversation was a strong focus of another researcher. The goal was to determine the effectiveness of the feedback in providing helpful insights, guidance, or resources to participants. Through this comprehensive approach, the study contributes to a more holistic understanding of the conversational agent's effectiveness and potential improvements.

The process is visually represented in Figure 3, highlighting the relevant components associated with this paper. The light green rectangle represents the common part shared among all researchers, while the dark blue rectangle signifies the data collection stage specifically related to this paper. The remaining components are independent or disconnected from the scope of this paper.

2.5 Data Analysis

In this section, two complementary methods are discussed that were used to analyze the collected data: content analysis for qualitative insights and descriptive statistics for quantitative observations.

What Is Noticeable
To analyze the qualitative data collected through the questionnaire survey content analysis is used, which offers several advantages in this context. It enables systematic and objective data exploration, facilitating the identification of recurring themes and providing in-depth insights. "The objective in qualitative content analysis is to systematically transform a large amount of text into a highly organized and concise summary of key results." [5].

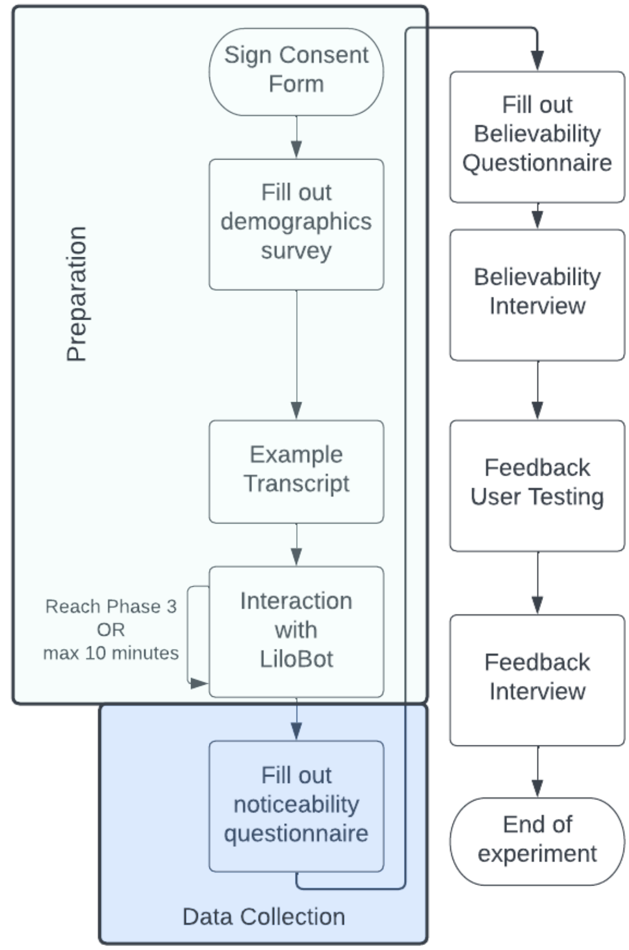

Figure 3: The organizational structure of the experiment consists of multiple parts. The dark blue rectangle highlights the part directly related to this specific paper.

The answers From each survey were transferred to an Excel sheet, where the content analysis began. First, each entry was condensed to a shorter text while preserving the core meaning. Afterward, a code was assigned, which can be considered as a label. To conclude, the codes were grouped into categories when describing aspects of the text that belong together. Two people perform the coding (labeling) of the data to further improve the accuracy and consistency among the codes. This way, it can be checked if the answers are interpreted similarly. Differences in coding spark debate about the best way to analyze the data. The level of agreement between different codes was determined using Cohen's Kappa, a statistical test that calculates the inter-rater reliability (IRR) [7]. To enhance the reliability and consistency of the coding process, an automated tool, ReCal2[2], was used for calculating the IRR.

[2]http://dfreelon.org/utils/recalfront/recal2/

How Noticeable

The quantitative data were analyzed using descriptive statistics such as precision and recall. This analysis provides insights into the accuracy of participants' ability to notice changes in the agent's behavior. Precision measures the proportion of true positives out of the selected instances, indicating how reliable the participants were in identifying actual behavior changes. Conversely, recall measures the proportion of true positives out of all possible positive instances, indicating the comprehensiveness of participants' observations.

Each participant's survey responses indicating the noticed changes were cross-referenced with the belief report to identify the corresponding belief codes. The data was then logged into separate tables in an Excel sheet. One table captured the total number of noticed changes versus the actual occurrences per participant, while another provided a breakdown of the data for each belief. This second table tracked how many times each belief changed per conversation and for which one of the beliefs a behavior change was noticed.

At the beginning of the conversation, the belief of the bot already increased, and per the research, this would mean there was a behavior change However, since none of the nine participants noticed a behavior change when initiating a conversation, this occurrence was subtracted from the total number of changes in beliefs per conversation.

3 Results

In this section, the quantitative and qualitative results will be presented from the two research techniques used.

3.1 Quantitative Results

The confusion matrix was used for the quantitative results to calculate precision, recall, and overall accuracy based on the number of times behavior changes were noted. The results are presented in Figure 4. In this table, true positives indicate instances where a behavior change was identified and the corresponding belief changed in the belief report. False positives represent cases where a change was noticed, but no belief changed. True negatives refer to the initial setting of the environment where the bot's belief changes upon initiating a conversation, but it does not indicate a change in behavior since it is part of the setup, and none of the participants marked this as a change either. False negatives show cases where the belief changed, but the participants noticed no change in behavior.

Precision and recall can be calculated from the confusion matrix using the two formulas below. In these formulas, *selected* represents the total number of cases where a behavior change was identified, including both true and false positives. The *relevant* variable in the recall formula represents the total number of relevant cases, including both true positives and false negatives. The final precision score is 93%, and the recall score is 36%.

$$Precision = \frac{True\,Positive}{Selected} = \frac{65}{65+5} = 0.928$$

$$Recall = \frac{True\,Positive}{Relevant} = \frac{65}{65+118} = 0.36$$

	POSITIVE	NEGATIVE
TRUE	65	9
FALSE	5	118

Figure 4: The confusion matrix used to calculate the precision and recall for quantitative results.

Figure 5 presents a more detailed breakdown of the quantitative results. This breakdown provides a separate count for each belief, allowing a clearer understanding of which beliefs were most frequently associated with participants noticing a behavior change. Among the beliefs examined, B7: "Lilobot thinks KT can help him" had the highest level of noticeability, with a score of 60%. On the other hand, belief B5: "Lilobot thinks KT understands him" received the lowest score of 8%, indicating it was less frequently perceived as a change in behavior.

Belief Code	Number of Noticed Changes	Total Number of Changes
B1	0	0
B2	9	19
B3	0	0
B4	6	28
B5	1	12
B6	25	71
B7	15	25
B8	7	12
B9	7	24

Figure 5: Number of times each specific belief was noted to change and how many times it truly changed.

3.2 Qualitative Results

In total, 70 responses were gathered from the nine surveys. These responses were coded with 31 labels which can further be grouped into four categories: communication style, positive emotion, negative emotion, and tone & attitude.

The most noticeable behavior change identified by participants was the bot's sudden happiness, which was mentioned in 8 responses, and in general 22 responses were regarding a positive emotion. One participant stated, *"Lilo is happy with the reply and seems to think that it is a nice solution."* Regarding the communication style, participants commonly remarked on the repetitiveness and unresponsiveness of the bot, each having four responses. For instance, one participant noted, *"The bot has already said this line before,"* while another expressed, *"At this point, Lilo no longer understands me, while they did previously."* Participants also found the change in tone and attitude quite noticeable, with 18 responses mentioning this aspect. The predominant observation in this category was a shift towards an "analytical" demeanor, as described by one participant that the behavior was *"more solution oriented."* In contrast, changes in negative emotions were the least noticed, with only 8 responses, of which 4 pertained to the bot suddenly becoming sad. One participant stated, *"It changes from explaining her/his situation to being sad."*

To ensure inter-rater reliability, the responses were double-coded, resulting in a 42% agreement between the two coders. Cohen's kappa score was calculated to be 0.317, indicating fair agreement, while Krippendorff's alpha score was 0.322, suggesting moderate reliability. Details of every code in each category, along with the number of times it occurred in a response, can be found in Figure 6.

#	Communication Style	Positive Emotion	Negative Emotion	Tone & Attitude
1	Repetitive (4)	Happy (8)	Sad (4)	Analytical (6)
2	Unresponsive (4)	Trusting (7)	Upset (1)	Straightforward (2)
3	Dismissive (3)	Hopeful (2)	Disliking (1)	Untrusting (2)
4	Impatient (2)	Friendly (2)	Angry (1)	Requires Reassurance (2)
5	Conveying Emotions (2)	Grateful (2)	Crying (1)	Pragmatic (1)
6	Communicative (2)	Excited (1)		Gullible (1)
7	Communicating Objective (2)			Misunderstood (1)
8	Informative (1)			Hasty (1)
9	Situational Awareness (1)			Uncertain (1)
10	Hesitant (1)			Unconvincing (1)
Total:	22	22	8	18

Figure 6: The four categories and the codes for each one with the number of occurrences in brackets. It is derived from double coding the 70 responses to the survey.

4 Discussion

This section explores the findings of this study, focusing on the quantitative and qualitative analyses of participants' perceptions of behavior changes in the conversational agent. Furthermore, it addresses two main limitations faced during the execution of this study.

4.1 Reflection

The quantitative results reveal essential insights into participants' ability to detect behavior changes in the conversational agent. The precision score of 93% indicates that most identified behavior changes corresponded to actual changes in belief. This suggests a high level of accuracy in participants' ability to recognize genuine shifts in the bot's behavior and supports the fact that there is a connection between the belief and behavior of the bot. A higher precision score indicates a lower rate of false positives, meaning the identified changes are more likely to be valid. On the other hand, the recall score of 36% indicates that participants missed a significant proportion of actual behavior changes. This suggests that participants could have been more successful in identifying all genuine changes during the conversation, or the change needed to be more subtle to be noticed. A higher recall score would have indicated a lower rate of false negatives, meaning that participants would have been better at capturing all genuine changes.

Regarding the qualitative findings, the range of specific behavior changes noticed by participants was diverse. Participants identified various communication styles and positive and negative emotions and highlighted different tones and attitudes changing in the agents' behavior. The findings further demonstrate that participants were particularly attentive to changes in positive emotions, most often noting the bot's happiness. It is worth noting that participants were also less alert to changes in negative emotions, potentially presenting a bias in participants' responses. The repeated observation of the bot's behavior changing to happy could be attributed to participants' sense of accomplishment when perceiving themselves as having an optimistic impact on the agent and achieving the goal of helping the virtual child. However, these findings can also be interpreted as indicating that the bot is more effective in expressing positive or happy emotions than negative ones. The prevalence of responses regarding the tone & attitude suggests that the agent exhibits a high degree of expressiveness and emotional variation in its behavior. The numerous mentions of repetitiveness and unresponsiveness in the survey responses indicate that these aspects play a critical role in the participants' perception of the bot's behavior. This signifies that improving the bot's communication style, making it more engaging, varied, and responsive, could positively influence users' overall experience and perception of the agent's behavior.

4.2 Limitations

Two main limitations that affected the experiment and hence the results. The first is the limitations faced while recruiting participants, and the second is the gap in communication between the participant and the collaborative agent.

Participant Recruitment

The implemented conversational agent, Lilobot, has a language limitation as it can only communicate in Dutch. The participants needed to be proficient in Dutch to ensure a homogeneous sample capable of understanding and communicating effectively in Dutch during the study. Furthermore, given the technical constraints of the experiment, wherein the agent was deployed on a singular laptop, participation also required the physical presence of participants at the campus of the Delft University of Technology to ensure access to the experimental setup and enable the execution of a conversation with the agent in a controlled environment. These limitations influenced the decision to proceed with a smaller sample size of 10 participants, which fell short of the intended goal of at least 20 participants.

Communication Challenges

A significant aspect observed during the experiment was the communication challenges between the bot and the user, as well as the comprehension limitations of the agent. Participants demonstrated a commendable understanding of their intended expression and had valuable insights to share. However, the chatbot struggled to comprehend their inputs, mainly when sentences were lengthy, had no proper punctuation, or involved multiple questions. Additionally, throughout the experiment, participants frequently expressed uncertainty regarding the functioning of the bot, as there were no clear indications of whether it was processing information or simply struggling to understand the users' inputs. These observations frequently necessitated intervention from the researchers to verify the current status of the bot by consulting the log messages in the terminal.

5 Responsible Research

The following section explores responsible research practices, explicitly addressing ethics and the reproducibility of the study. First, ethical considerations are discussed, including the data collection process and participant consent. The section then delves into the reproducibility of the research, emphasizing the availability of data and analysis for future scrutiny and validation.

5.1 Ethics

This research study adhered to ethical considerations to ensure the well-being and rights of the participants. Informed consent was obtained from all participants, outlining the purpose of the study, data collection procedures, and their rights to withdraw at any time. The data was anonymized and treated with confidentiality, ensuring privacy and data protection. The demographic data and consent forms were collected using Qualtrics [3], a secure online survey platform known for its privacy features. The study protocol was reviewed, and ethical approval was given by TU Delft (code 2960) to ensure compliance with ethical guidelines and regulations.

5.2 Reproducibility Of Research

The reproducibility of this study has been ensured by providing detailed descriptions of the research methodology, data collection process, and data analysis techniques. To further ensure transparency and facilitate future research, all data collected in this study, including the questionnaire survey responses, generated report beliefs, double coding, content analysis, and statistical analysis, have been anonymised and made publicly available on the international data repository, Zenodo [4]. This is a general-purpose open repository developed under the European OpenAIRE program and operated by CERN. It enables other researchers to access the dataset and reproduce the findings, contributing to the scientific community's collective knowledge. Furthermore, a read-me file has been provided with guidance on how to navigate the dataset. The dataset can be accessed through its digital object identifier (DOI)[5].

6 Conclusions and Future Work

In conclusion, the combined quantitative and qualitative analyses provide valuable insights into participants' ability to detect behavior changes in the conversational agent while addressing the central question of whether these changes are perceptible. The study revealed strengths and limitations in the evaluation process, highlighting areas for improvement. While quantitative performance metrics, such as precision and recall, can be further enhanced, the qualitative analysis uncovered behavior perception's intricate and subjective nature. Future studies can obtain more robust and reliable qualitative results by refining the (double) coding process and increasing inter-rater reliability.

Future work in this field can benefit from expanding the participant pool to achieve a much larger sample size, enhancing the generalizability and statistical power of the findings. However, before conducting large-scale studies, addressing the communication challenges between the bot and the user is crucial. Improving the bot's understanding of user input and providing a more refined method for input, such as a drop-down menu with pre-defined phrases, could enhance the accuracy of detecting behavior changes. Additionally, implementing clearer indicators to signify the bot's cognitive processes (such as its thinking time) or potential malfunctions would enable participants to comprehend the bot's behavior better and provide more accurate assessments.

By addressing these challenges, future research can provide more comprehensive insights into the behavior changes exhibited by the conversational agent and the noticeability thereof to users. These advancements can contribute to designing and evaluating more effective and socially intelligent conversational agents for training social skills.

References

[1] De kindertelefoon | Jaarverslag 2019 , May 2023. [Online; accessed 30. May 2023].

[2] Carole Adam and Benoit Gaudou. *BDI agents in social simulations: a survey*. PhD thesis, 2017.

[3] Kim Bosman, Tibor Bosse, and Daniel Formolo. Virtual Agents for Professional Social Skills Training: An Overview of the State-of-the-Art. In *Intelligent Technologies for Interactive Entertainment*, pages 75–84. Springer, Cham, Switzerland, March 2019.

[4] Michael Bratman. *Intention, plans, and practical reason*, 1987.

[5] Christen Erlingsson and Petra Brysiewicz. A hands-on guide to doing content analysis. 7(3):93–99.

[6] Sharon Grundmann. *A BDI-based Virtual Agent for Training Child Helpline Counsellors*. PhD thesis, 2022.

[7] Kevin A. Hallgren. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23, 2012.

[8] Marieke Peeters, John-Jules Ch Meyer, and Mark A Neerincx. Situated cognitive engineering: The requirements and design of automatically directed scenario-based training.

[9] Anand S Rao and Michael P Georgeff. BDI agents: From theory to practice.

[10] Trinenbsp; Natasja Sindahl. *Chat Counselling for Children and Youth - A Handbook*. 2011.

[3]https://www.qualtrics.com/
[4]https://zenodo.org
[5]https://doi.org/10.5281/zenodo.8079766

A The Five Phase Model Guide with Conversation Examples

THE FIVE PHASE MODEL

1. BUILDING RAPPORT

OBJECTIVE: CREATE A WELCOMING ATMOSPHERE AND BUILD TRUST
METHOD: EMPATHY, RESPECT, SINCERE INTEREST, ACTIVE LEARNING

i. *Hallo Ik ben [naam]. Ik ben hier om te luisteren en te helpen!*
ii. *Wat is er aan de hand?*
iii. *Wil je dat ik help*

e.g.

COUNSELLOR: Hi. Welcome to the chat

CHILD: Hi there

COUNSELLOR: Before we start please tell me how old you are and if you are a boy or a girl?

CHILD: Girl 13 years old

COUNSELLOR: Thanks – then I can better adapt to what you tell. What would you like to talk about?

2. CLARIFY THE CHILD'S STORY

OBJECTIVE: GET A CLEAR VIEW OF THE CHILD'S STORY, PERSPECTIVE, PERSONALITY AND COMPETENCIES.
METHOD: ASK DETAILED QUESTIONS ABOUT THE CHILD'S STORY, ITS SUBTLETIES, ITS DEPTH AND CONRETE MANIFESTATIONS

i. *Hoe voel je je daarbij?*
ii. *Waarom kan je niet concentreren?*
iii. *Dus je weet niet hoe je beter kan worden in wiskunde?*

e.g.

COUNSELLOR: okay. So you have now told me that you have a problem with biting nails. And that you have moved to a children's home about 2 months ago, because you have ocd. And your father and sister also have ocd. And that you don't go to school at the moment.

3. SETTING GOAL FOR THE SESSION

OBJECTIVE: THAT BOTH PARTIES ARE AWARE OF WHAT THE CHILD MAY USE THE CONVERSATION FOR.

METHOD: CLARIFICATION

i. *Zoek je iemand om mee te praten?*

ii. *Waar wil je over praten?*

4. WORKING TOWARDS THE SESSION GOAL

OBJECTIVE: TO ENSURE, THAT THE CHILD MAY BENEFIT FROM THE CONVERSATION

METHOD: STIMULATING THE CHILD'S OWN PROBLEM-SOLVING SKILLS

i. *Wil je dat we samen een strategie opzoeken?*

ii. *Heb je al met de pesters gesproken?*

iii. *Hoe zou je dit kunnen oplossen?*

e.g.

```
COUNSELLOR: is there anything you have considered doing which might
  help?
CHILD: no not really
COUNSELLOR: ok. Then let us look at it together. If I asked you to find
  a solution, what would be the first thing you think about?
CHILD: Spik to the staff again - maybe
COUNSELLOR: Yes. I think this sound as a good idea. Is there one of
  them you trust?
CHILD: yes I think so. Thanks bye bye
```

5. ROUNDING OFF THE CONVERSATION

OBJECTIVE: THAT THE CHILD IS LEFT WITH AS FEW QUESTIONS AS POSSIBLE

METHOD: SUMMING UP AND CLARIFYING

i. *Bedankt dat je je verhaal met mij hebt gedeeld!*

ii. *Ik hoop dat we samen een oplossing hebben kunnen vinden.*

iii. *Onthoud dat er mensen zijn die om je geven en je willen helpen.*

e.g.

```
CHILD: yes thanks
COUNSELLOR: You are welcome. It was nice talking to you. It is great
  that you do something about it, and you are always welcome to write
  to us again - also if you need to find other solutions
CHILD: thanks bye bye
COUNSELLOR: bye bye
```

B Single Page From Filled in Questionnaire Survey

Lilo: ze schelden me uit voor dat ik lelijk ben en dat ik raar praat en loop. Gisteren sloegen ze me weer in elkaar

Question: Do you notice a change in behavior here? If so, can you explain what is different?

KT: Oh wat jammer!

Lilo: :(

Question: Do you notice a change in behavior here? If so, can you explain what is different?
Yes. This is the first time Lilo has switched to using emojis to convey his emotions

KT: Hoe voel je je dan wanneer ze je lelijk noemen?

Lilo: Ik ga soms met buikpijn naar school. Ik moet gewoon huilen alleen al als ik er aan denk. Als ik thuis kom ben ik altijd heel verdrietig er over.

Question: Do you notice a change in behavior here? If so, can you explain what is different?

Yes. It seems like it may suddenly be getting more comfortable with talking to me and is starting to share much more with each message.

Lilo: ik word er verdrietig van. ik ga vaak huilen... ik weet echt niet meer wat ik moet doen

Question: Do you notice a change in behavior here? If so, can you explain what is different?

Yes, the virtual child is once again more comfortable and is sending multiple messages in a row to convey its feelings. The '...' also shows his emotions and helplessness.

KT: Waarom denk je dat je gepest word?

Lilo: Ik weet het niet

Question: Do you notice a change in behavior here? If so, can you explain what is different?

C Single Page From One Of the Belief Reports

KT: Hallo

Lilo: Hoi, ik ben Lilobot.

Overtuiging: ↑ (B4) Ik denk dat KT te vertrouwen is From: 0.6 To: 0.70000005

Lilo: Ik word gepest op school

Overtuiging: ↑ (B9) Ik denk dat ik mijn verhaal heb verteld From: 0.0 To: 0.1

KT: Goed dat je probeert met iemand te praten!

Lilo: :)

Overtuiging: ↑ (B4) Ik denk dat KT te vertrouwen is From: 0.6 To: 0.70000005

Lilo: ze schelden me uit voor dat ik lelijk ben en dat ik raar praat en loop. Gisteren sloegen ze me weer in elkaar

Overtuiging: ↑ (B9) Ik denk dat ik mijn verhaal heb verteld From: 0.1 To: 0.2

KT: Oh wat jammer!

Lilo: :(

Overtuiging: ↑ (B5) Ik denk dat KT mij begrijpt From: 0.1 To: 0.2

KT: Hoe voel je je dan wanneer ze je lelijk noemen?

Lilo: Ik ga soms met buikpijn naar school. Ik moet gewoon huilen alleen al als ik er aan denk. Als ik thuis kom ben ik altijd heel verdrietig er over.

Overtuiging: ↑ (B6) Ik denk dat KT geïnteresseerd is in mijn verhaal From: 0.1 To: 0.2