Evaluating the Believability of the Lilobot Conversational Agent

A Bachelor Thesis

Vladimir Makarov

Supervisors: Willem-Paul Brinkman, Mohammed Al Owayyed

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 23, 2023

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

Abstract

The aim of this research is to evaluate the believability of Lilobot, a conversational agent meant to act as a virtual child for training helpline workers. Numerous aspects of believability are explored by means of a user study involving a questionnaire and interview with 10 participants. Questionnaire results indicate that improvement to the chatbot's believability is likely necessary.

The findings from the interviews are that the use of emoticons and acknowledging the context of the application raise believability, while unresponsiveness and repeated utterances lower it. While Lilobot did express valid and real emotions, study participants suggested improving the appropriateness of its reactions and expanding its vocabulary.

1 Introduction

Chatbots are a powerful tool for conveying information to humans in a personalized but controlled manner. One emerging use of chatbots is communication skills training [1]. Such training is necessary for many human endeavors, such as training helpline agents. To learn how to communicate in those scenarios, a caller would be impersonated by an experienced colleague and a trainee would interact with them. With the help of conversational agents, users (trainees) can instead interact with a chatbot that acts as the other party (i.e. a person texting the helpline). This aims to simulate the regular training process in a controlled setting.

This simulated environment is meant to help the user gain communication skills and avoid the need of using another human for training purposes. This training also greatly reduces risks for trainees, as there is no possibility to put another human in danger (e.g. of hurting their feelings). However, for the trainee to take the training seriously it needs to have an element of realism, such that the human believes that the conversation they are having could occur in real life.

This paper is concerned with the interaction between inexperienced trainees and an agent-based simulation of a virtual child (called Lilobot), that was bullied at school. The original work by Sharon Grundmann examines the usability and usefulness of the conversational agent [2], but did not measure the believability of the agent.

Previous research establishes the importance of assessing the believability of a conversational agent [3] and introduces relevant sub-topics and sub-questions. Past studies have also shown that text-based chatbots can be perceived as human-like based on their written behaviour [4]. Believability is especially important for this agent since it is trying to imitate the real-world behaviour of a child.

This leads to the following research question: How believable do trainees find the interaction with the Lilobot conversational agent?

The main contribution of this research is to evaluate the believability of Lilobot in acting as a bullied child. Through a user study, we aim to measure users' perceptions of the agent's believability and identify factors that may impact believability in this context. First-hand experience with the chatbot is valuable for the evaluation as it lets users form an opinion about the believability of the agent. But, second-hand experience (through viewing a transcript of a conversation) is more technically feasible in this specific scenario and allows for a bigger pool of study participants. Furthermore, the accuracy of the believability assessment can be positively influenced by using second-hand experience [5], as it helps the assessor concentrate on the believability of the agent.

In addition to that, being more removed from direct interaction eliminates the possibility of ”negative disinhibition” from occurring in a conversation. This is beneficial to all parties, as ”whenever agents attempt to assert themselves or to claim for themselves certain human rights and privileges,” users commonly respond with negative and offensive messages [6, p. 31].

Lilobot uses the beliefs-desirses-intentions (BDI) model to human-like behaviour [2]. Explaining the inner workings of the chatbot to the users in a way that is easy to understand could improve the experience of the interaction and the user's trust in the chatbot.

Although in this context the agent's main purpose is to imitate a human, there is no way for it to escape the connotations and treatment prescribed to a computer program. Such a reaction would definitely introduce a bias into the responses of our study participants, it is therefore helpful to avoid it.

The paper is organized as follows. Section 2 describes the methodology of our user study. Section 3 presents our results and findings. A discussion of the results in Section 4 is followed by Section 5 on responsible research. Finally, Section 6 concludes the paper and outlines directions for future work.

2 Method

This section will elaborate on the experimental procedure in this study. First, the demographics of the participants will be examined, then the materials used in the study will be described. Lastly the procedure will be discussed, including the design and organization of the study, and the evaluation method for collecting and processing the data.

2.1 Participants

The participants were recruited on site at the TU Delft university. The only requirement for a participant is sufficient knowledge of the Dutch language to understand Lilobot's conversations. We performed the study with 10 people, with 9 identifying as male and one as female. All 10 were in the 18-24 age range. All except one have had experience with chatbots, with 3 participants reporting using them often (more than 10 times a month), 4 saying they use them sometimes (2-10 times a month) and 2 reporting rare usage (once a month or less). One participant did not know Dutch, so we acted as a translator between them and Lilobot. Their results were not excluded as they still provided valuable feedback in the believability questionnaire and interview.

2.2 Materials

The materials used in this study involve a questionnaire and an interview to collect data. These are used to measure the different aspects of believability that we identified as applicable for Lilobot, so they will be elaborated on in the measures section. Alongside that, training material in the form of example transcripts is used to prepare participants for their interaction with the chatbot.

Example Transcript

Viewing example transcripts has the potential to demonstrate the five-phase model [7], which is the communication protocol that the children helpline uses to train counselors when interacting with children [2].

While demonstrating the 5 phase model and ideal scenarios of interaction is helpful to the participant, as it demonstrates that e.g. short sentences should be used when communicating with Lilobot, it also risks 'over-fitting' the participants' interactions with Lilobot. They might resort to using only the phrases we provided in their own interactions, rendering the evaluation results unusable. Therefore, a conversation transcript was created that is not related to the topic of Lilobot's conversations (bullying), and contains examples of questions for each of the 5 phases. Examples were taken from [7]. The example transcript is shown in Appendix C.

2.3 Measures

Believability Questionnaire

A questionnaire survey was used to gather quantitative data. The survey consists of questions presented by S. Fitrianie et al. in the artificial-social-agent questionnaire [8]. This questionnaire uses the 19 constructs found in [3], turning each into a set of questions. The first of the constructs is believability, which applies to this study. The construct they present is divided into 5 sub-constructs. We found that only 2 of the constructs could be applied to our topic, as Lilobot has no appearance beyond a chat window, therefore asking participants about it wouldn't yield meaningful results. Only the following 2 constructs will be used in our survey: "Human-Like Behaviour" and "Natural Behaviour". These contain 5 and 3 questions respectively, with answers provided on a 7 point Likert scale. The full questionnaire is shown as Figure 3 in Appendix B.

Believability Interview

An interview was used to gather qualitative data. The interview questions are shown in Appendix A. Since the believability of an agent depends on their ethos[1] [6], it is worthwhile to try to find what aspects of Lilobot's ethos are deemed believable or unbelievable by study participants.

Posing questions about what aspects of Lilobot's personality influenced believability can show what parts of the interaction are important for believability.

Asking how the conversation would differ if instead of Lilobot the participants were talking to a real child lets the participants reflect on the ethos they constructed for Lilobot and give insight into how Lilobot can shape the perceived ethos of the user to better fit its role. Questions 3 and 4 are follow-up questions to 2, exploring the different aspects of the participants' interaction with the chatbot.

The last question is concerned with the context of the chatbot. It is important to consider that Lilobot was created for a very specific use case and evaluating its believability in this context is helpful. The question used the "Agent's Appearance Suitability" section from the artificial-social-agent questionnaire [8] as inspiration, changing the word appearance to behaviour.

2.4 Procedure

Aside from data collection methods and experiment design, the overall study structure involving collaboration with other researchers, is described in this section.

Experiment Design

At the beginning of the experiment, the participants' consent and demographics data are collected. Then, the participants are prepared for their interaction with the chatbot by reading an example transcript, followed by a real-time interaction with Lilobot. The final structure of the experiment can be seen in Figure 1.

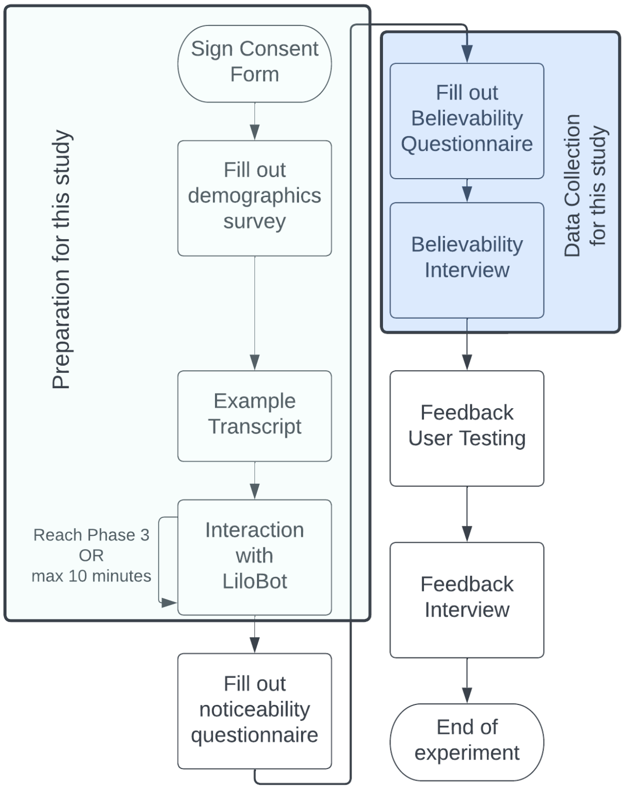

Preparation for this study
Sign Consent Form
Fill out demographics survey
Example Transcript
Interaction with LiloBot
Reach Phase 3 OR max 10 minutes
Fill out noticeability questionnaire
Data Collection for this study
Fill out Believability Questionnaire
Believability Interview
Feedback User Testing
Feedback Interview
End of experiment

Figure 1: The organizational structure of the experiment consists of multiple parts. The light green rectangle highlights the parts that relate to preparing participants and collecting demographics data for this study. The blue rectangle highlights the parts which are directly related to gathering the data for this study. The non-highlighted parts are entirely performed by other researchers for other studies, therefore they do not directly concern this study.

The study was performed in collaboration with other researchers also studying the evaluation of different aspects of

[1] This term has many meanings in different contexts, but can be crudely summarized as the reputation of a communicator.

Lilobot. Namely, the noticeability of behavioural changes and the usability of feedback provided at the end of the interaction. Much thought was given to the order in which the participants would answer the different questionnaires and interviews. The noticeability assessment precedes our study, as the participants need a fresh memory of their conversation with Lilobot to assess the precise moments when the chatbot changes its behaviour. The feedback assessment is last as it reveals a lot about the inner workings of the chatbot and may influence the other 2 studies by introducing biases to the participants' answers.

Conversation with Lilobot

To evaluate the believability of the Lilobot conversational agent, the study participants need experience with the agent, which they will then reflect upon. After viewing the example transcripts described in 2.2, the study participant engage in a real-time conversation with Lilobot. Due to technical limitations, Lilobot cannot recognize multiple intents in the user's prompts [2]. This forces users to write single sentence prompts, which have to be similar to ones which the chatbot was trained on. Studying example transcripts with plenty of good and bad examples can help the participant have a meaningful conversation with Lilobot.

The conversation can end in a number of different ways, with the chatbot leaving the conversation after being disappointed or with the intent to talk with its teacher being the most common (the latter being the desired outcome). As seen in Figure 1, if the conversation ended early without reaching phase 3 of the 5-phase model, then the participant had to restart the conversation. This lasted for a maximum of 10 minutes due to time limitations.

A question that was considered during the design of the study is whether the nature of a participant's interaction with Lilobot plays a role in their assessment of the chatbot's believability. The participant's view of Lilobot's believability may be formed from 3 sources: observation of the example transcripts, direct interaction with the chatbot, and any preconceptions that they may have had before participating in the study. The proportion and importance of each is unknown and likely depends on the specific person. However, it is a risk that with a (un)favorable choice of words and actions, the participant might experience a best-case (or worst-case) scenario in their conversation with Lilobot which might adversely affect their assessment. Out of the three sources, only the observation the example transcripts is controlled and stays constant between participants.

Data analysis

The raw data was collected in the following formats. The questionnaire (as seen in Figure 3 in the Appendix) has 8 questions which are divided into 2 groups. The answer is in the format of a 7 point Likert[2] scale, which was mapped to the numbers -3 to 3, from strong disagreement to strong agreement respectively.

Interview transcripts were written during the interview and any identifying information was removed. A content analysis performed on the interview transcripts according to the steps outlined in [9]. Meaning units were extracted from the interview transcripts, followed by making codes and categories. The codes were generated inductively. Double coding was utilised on 12 meaning units to assess and strengthen the reliability of our coding scheme. Two double coders were employed, both are computer science students and researchers working on other parts of the experiment shown in Figure 1. We found an average pairwise Cohen's κ of 0.396, which indicates fair agreement [10]. The codes had a Krippendorff's Alpha value of 0.395 . Table 1 shows all the categories that assembled the meaning units into. It also shows the number of quotes each category contains.

Category	Number of Quotes
content	66
content structure	13
role context	12
technological limitations	2
medium	1

Table 1: Showing all the categories and the number of quotes (meaning units) assigned to each category.

3 Results

3.1 Questionnaire results

The results of the artificial-social-agent questionnaire [8] can be seen in Figure 2. The mean value for the Human-Like Behaviour construct is -0.6 (SD = 1.65), and the mean for Natural Behaviour is -0.87 (SD = 1.66). Both map to being slightly below the 'Neither Agree nor Disagree' Choice on the 7 point Likert scale, leaning towards disagreement. As can be seen by the error bars representing the standard deviation, there is a wide range of opinions in the responses. The 'Lilobot behaves like a real person' metric is the most extreme out of the 8, favoring the disagreement side, but the neutral response still falls within the standard deviation.

3.2 Interview Results

The findings of the interviews were in places similarly contradictory. Six participants made comments on Lilobot's emotions being valid and real, with one participant (P3) saying "she explained the story right away without me asking is believable because it shows that Lilobot is looking for a solution to their problem. She had a reason to reach out". However, four participants also thought the opposite, with one participant (P6) remarking "It's not like a child at all, basically. Like it's way too robotic". There was a clear sentiment that Lilobot was not believable at some points.

Conversation Pace and Progression

The pace and speed of progression of the conversation were also difficult and contentious topics. Eight participants made remarks related to the conversation progressing too fast and the chatbot being too prepared to tell them about itself. Participant 2 recommended that for more believability, "Maybe

[2] The Likert scale is commonly used in questionnaires to measure the scale of agreement or disagreement with a statement. Includes a neutral option.

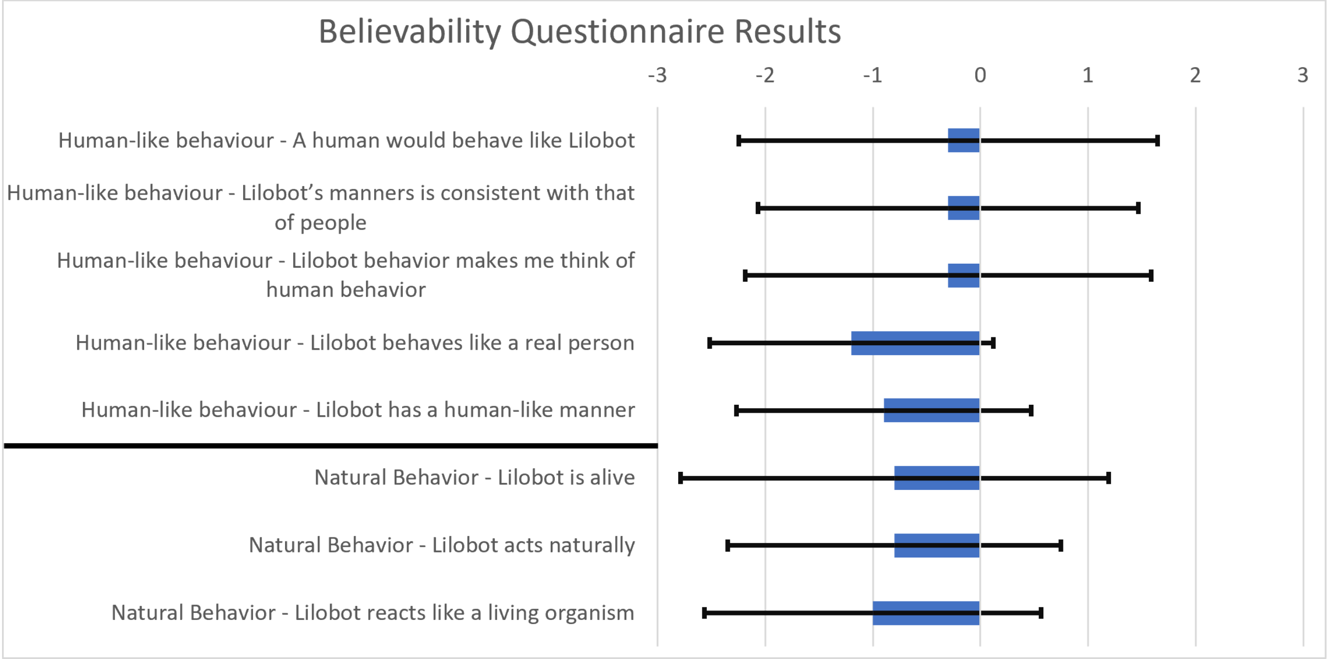


Figure 2: Results of the artificial-social-agent questionnaire [8]. The first 5 metrics are parts of the Human-Like Behaviour construct, and the latter 3 are part of the Natural Behaviour Construct. The values are mapped from the 7-point Likert scale, with -3 representing extreme disagreement and 3 representing extreme agreement. The blue bars indicate the mean value of the metric and the black error bars indicate the standard deviation.

[Lilobot's responses] wouldn't be that detailed, maybe I'd have to ask it more questions before it started giving me the amount of detail that I needed".

This sentiment, was however reversed when it came to the task of reaching past phase 2 of the 5-phase model, as when the trust requirements of the robot are not met (it needs to tell enough about its situation to move past a threshold of trust), the participant is seemingly barred from entering the latter stages and the conversation grinds to a halt. Participant 4 experienced this very strongly, which resulted in them saying both that "I didn't really believe the story because it went too slow, as in it wasn't urgent." and "I would expect it to more slowly, gradually, go into the problem". These opposing statements uttered within 1 minute of each other can signify that there needs to be a delicate balance in the actions of the chatbot, preventing it from oversharing too quickly, but letting frustrated users progress to further phases of the 5-phase model in their conversation, for training purposes.

Complaints and Improvements

Lilobot's repeated utterances of the same question (e.g. "when will you call my school?") affected the believability negatively. Three participants noticed that occurrence, for example, P3 said "She kept repeating a question that I already answered. Lilobot asked it three times. That makes it a bit less believable". Furthermore, the chatbot often reacted incorrectly to the participants' prompts. Six participants observed inappropriate reactions, with P5 stating "I think when I said 'tell what you told to me to your teacher', the reaction the was just a sad face, that was weird". Another identified issue was unresponsiveness or failure to respond to messages, which was mentioned by three participants. Unresponsiveness can likely be fixed by expanding the dictionary of phrases that the chatbot can understand.

Two other technological limitations that participants noticed were that the chatbot had no memory of previous prompts, as P8 pointed out that it's "kind of limiting how you could respond yourself. Probably can't refer to previous comments". Also, the wording of the users' prompts had to match certain criteria, as P9 stated that "The responses had to be formatted in a very specific way". These software limitations affected the believability negatively, as in a real-life conversation with a child varied sentence structures would be understood.

Three participants also noticed fast and abrupt changes of topics or emotions. They could emerge from a lack of consistency in Lilobot's emotions, as P1 states "A real child would be a bit more consistent with their emotions and you would be able to maybe establish more of a connection with the kid", and P8 stated that "It was a quick transition from one prompt to another, that felt a little bit unnatural".

A complaint about the chatbot that was expressed by three participants was the lack of a backstory, or personal information about the character that Lilobot is acting as. Participant 7 said that when speaking to a real child, they "would feel more empathy towards the child, because with the bot you cannot know their gender and age. Because you cannot know that, it becomes less human".

Positive sentiments

On the other hand, the use of emoticons such as ":)" in Lilobot's vocabulary produced a notable positive effect on the believability. Five participants remarked clearly that it helped the chatbot seem more realistic. In addition, in answer to the last interview question (shown in Appendix A), Six participants contextualised the shortcomings of the chatbot or said that it was suitable for its intended use of teaching the 5-phase model. Participant 5 explained that ”you don’t really have to have a perfectly working, completely human child. Because every child is different, and this does really train you for the [5-phase model]”.

4 Discussion

In this section, we assess our findings and contributions by comparing them to other researchers and giving our interpretation of the results. The issues and limitations in our methodology are also examined, along with recommendations for future research.

4.1 Artificial-Social-Agent Questionnaire

To put our results into a general perspective, we can compare our values for the Human-Like Behaviour and Natural Behaviour constructs to the ones that were found by researchers that developed the questionnaire in [11]. Compared to our values of -0.6, and -0.87 respectively, some similar agents achieved similar results. Siri, a virtual assistant developed by Apple scored -0.28 and -0.81 respectively. While Lilobot scores lower in both constructs, it is also unexpectedly close to a chatbot made by one of the biggest companies in the World. In addition to that, Siri can communicate using sound, which could give it an advantage over Lilobot. Poppy, a virtual human from SEMAINE [12] scores 0.51 and -0.38 respectively. A large difference between Poppy and Lilobot is that Poppy has a virtual avatar, and can therefore communicate both with sound and visual body language. It scores far better in the human-like behaviour construct, but the natural behaviour construct value is still similar to Lilobot’s.

4.2 Repeated Utterances

As discussed in the Results section earlier, a common complaint about the chatbot was that it repeated questions multiple times. There are several possible causes and solutions for this behaviour. One possible cause is that there is a bug in the software of Lilobot causing it to print prompts multiple times, or it could be intended behaviour for repeatedly trying to get an answer to an important question. A possible solution could be to prefix a repeated sentence with a fitting remark, for example ”What about ...”. This would seem more believable, as there is a change to the prompt each time, making it less repetitive. Another possible cause is that the participant had given a reply, but it wasn’t understood by the chatbot, so improving the dictionary of inputs and allowing for multiple sentiments to be processed from one prompt is a possible solution[2].

4.3 Unresponsiveness

A lack of response to the participant’s prompt caused dissatisfaction and annoyance in participants. P4 said that ”It would be acceptable that the kid would tell you more, or say random stuff in between, but the bot just gave no feedback”. This reaction leads to several ways to alleviate the inevitable situation when the chatbot receives a message it cannot understand. Rather than providing no response at all (which is what the chatbot does currently), a polite acknowledgement of failure should be provided[13]. Since the chatbot is imitating a child, the inexperience and lack of knowledge attributed to a young human can be used as a means for graceful failure. As mentioned previously, the number of occurrences of unresponsiveness can be reduced by adding to Lilobot’s understanding.

4.4 Lack of backstory

While it is natural that the chatbot cannot provide any personal details outside the bullying scenario since the character is not based on a single real person, some participants suggested that it would improve the believability of the conversation to provide some information on the gender and age of the chatbot. This remark may have been caused in part due to the example transcript containing that information (but from another scenario). However, implementing this idea could open avenues to stereotyping based on personal characteristics or decrease the generalisability of the training to the diversity of real-life conversations. In addition to that, as mentioned before, negative disinhibition might occur if too many human characteristics are assumed by the chatbot [6]. Therefore, from our perspective, the addition of personal details remains an open question.

4.5 Limitations

The first and foremost limitation of this study is the software of of Lilobot. We observed that it was slow to run and prone to errors when non-ideal inputs were given. In addition to that, since it was not hosted online we had to run the chatbot locally from one computer. This restricted us to conducting the survey in-person, whereas an online survey could have reached more people and given a larger quantity of results.

Secondly even though the intended usage of the software was to train children’s helpline workers [2], we only had access to the general public for recruiting participants. Therefore, we had to spend more time acquainting the participants with the 5 phase model and were forced to avoid particularly serious and negative topics to show for training. This was an issue because the handbook that introduces the 5 phase model uses a very graphic and serious conversation as the only example transcript [7]. Since people unaccustomed to helpline problems might be more sensitive to such topics, we were advised by our supervisor to not use this material for fear of disappointing or traumatizing our study participants. Additionally, the diversity of our recruited participants is severely lacking. Only one out of the 10 identified as Female and all were in the same age-range (18-24) this may have affected our results, as we do not know the demographics of the target user group of the chatbot and therefore cannot determine if our results apply to it.

Thirdly, given more time to explain the 5 phase model and the main principles of conducting a helpline conversation to the participants, our results could potentially be more meaningful. Unfortunately as we were conducting the experiment in collaboration with other researchers doing their own studies, and the general time constraints of the project, the time to introduce the topic and educate the participants was limited to around 10 minutes. This is in stark contrast to the intended users of Lilobot, who are children's helpline workers undergoing a training lasting several weeks [14].

After receiving complaints about the beliefs of the robot from some participants, we found that the database containing the beliefs of the chatbot was not being reset between tries with different participants. This may have affected the results in one of the first 3 runs of the experiment by altering the behaviour of the chatbot. After we found this out, we started resetting the database values after every run of the experiment. We did not, however, discard the data up to that point because apart from aggravating some negative emotions from the participants, the responses were similar to the proceeding participants.

5 Responsible Research

The major ethical concerns in this study are collecting user-identifiable data. Apart from using their name to sign the consent form (those names are not shared with anyone), participants provided their age, gender and experience with chatbots in a demographics questionnaire. Of these, only age and gender are potentially identifiable characteristics, and will only be shared in the form of statistics for the whole population, with no specific response being linked directly to the participant. Any personal information is also censored from the interview transcripts. The data collected for this study is published in the 4TU repository[3]. To collect demographics data and perform the questionnaire, the survey tool Qualtrics[4] was employed, known for its compliance with privacy laws (GDPR). The qualitative analysis performed in this study was verified by the use of double coding with another researcher.

Reproducing our methods is fully possible. We provide the full questionnaire and interview questions used in the study, and apart from the recruiting the same participants, every part of the study can be repeated. The software used (Lilobot) is taken from [2]. The research also received ethical approval from TU Delft, reference number 2960.

6 Conclusions and Future Work

The results of the ASA questionnaire were inconclusive. The average answer leaned towards disagreeing with the believability of the chatbot, however the range of opinions was large. This leads to a conclusion that an improvement in believability is likely necessary for the chatbot. In the interviews, the participants noted that the emotions that Lilobot was talking about were real and valid, though the structure of the responses decreased the overall believability. This included Lilobot repeating questions multiple times or being unresponsive. The use of emoticons and acknowledging the context of the chatbot have a positive effect on the believability. Believability can also be improved with better reactions to the participants' messages, which can be achieved by broadening the vocabulary of the chatbot. An interesting area for further research is extending the functionality of the chatbot to react to new and unexpected prompts, for example by the use of Large Language Models[15], though there are potential ethical issues [16].

References

[1] Kim Bosman, Tibor Bosse, and Daniel Formolo. Virtual agents for professional social skills training: An overview of the state-of-the-art. In Paulo Cortez, Luís Magalhães, Pedro Branco, Carlos Filipe Portela, and Telmo Adão, editors, *Intelligent Technologies for Interactive Entertainment*, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, page 75–84, Cham, 2019. Springer International Publishing. ISBN 978-3-030-16447-8. doi: 10.1007/978-3-030-16447-8_8.

[2] Sharon Grundmann. A bdi-based virtual agent for training child helpline counsellors. 2022. URL https://repository.tudelft.nl/islandora/object/uuid%3Af04f8f0b-9ab9-4f1c-a19c-43b164d45cce.

[3] S. Fitrianie, M. Bruijnes, Deborah Richards, Andrea Bönsch, and W. P. Brinkman. The 19 Unifying Questionnaire Constructs of Artificial Social Agents: An IVA Community Analysis. *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents, IVA 2020*, 2020. doi: 10.1145/3383652.3423873. URL https://repository.tudelft.nl/islandora/object/uuid%3A521411a1-579c-4054-878d-31ba3815b628. Publisher: Association for Computing Machinery (ACM).

[4] Amon Rapp, Lorenzo Curti, and Arianna Boldi. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies*, 151:102630, Jul 2021. ISSN 10715819. doi: 10.1016/j.ijhcs.2021.102630.

[5] *Believable Bots: Can Computers Play Like People?* Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-32322-5. doi: 10.1007/978-3-642-32323-2. URL https://link.springer.com/10.1007/978-3-642-32323-2.

[6] Sheryl Brahnam. Building character for artificial conversational agents: Ethos, ethics, believability, and credibility. *PsychNology Journal*, 7:9–47, 2009. ISSN 1720-7525.

[7] Trine Natasja Sindahl. *Chat Counselling for Children and Youth: A Handbook*. Børns Vilkår, 2011. Google-Books-ID: yS8GrgEACAAJ.

[8] Siska Fitrianie, Merijn Bruijnes, Fengxiang Li, Amal Abdulrahman, and Willem-Paul Brinkman. The artificial-social-agent questionnaire: establishing the long and short questionnaire versions. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, IVA '22, page 1–8, New York, NY, USA,

[3] https://doi.org/10.4121/d24f832f-748c-4ac7-ba7b-879637c6c64d.v1

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

Sep 2022. Association for Computing Machinery. ISBN 978-1-4503-9248-8. doi: 10.1145/3514197.3549612. URL https://dl.acm.org/doi/10.1145/3514197.3549612.

[9] Christen Erlingsson and Petra Brysiewicz. A hands-on guide to doing content analysis. *African Journal of Emergency Medicine*, 7(3):93–99, Sep 2017. ISSN 2211419X. doi: 10.1016/j.afjem.2017.08.001.

[10] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. ISSN 0006-341X. doi: 10.2307/2529310.

[11] Siska Fitrianie, Merijn Bruijnes, Fengxiang Li, Amal Abdulrahman, and Willem-Paul Brinkman. Data and analysis underlying the research into the artificial-social-agent questionnaire: Establishing the long and short questionnaire versions, Jul 2022. URL https://data.4tu.nl/articles/_/19758436/2. DOI: 10.4121/19758436.V2.

[12] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, Jan 2012. ISSN 1949-3045. doi: 10.1109/T-AFFC.2011.20.

[13] Hans van Dam. Error handling: tips to make your chatbot more helpful, natural and persuasive, Mar 2021. URL https://blog.conversationdesigninstitute.com/error-handling-tips-to-make-your-chatbot-more-helpful-natural-and-persuasive. Accessed: 8/6/2023.

[14] Kindertelefoon homepage. URL https://www.kindertelefoon.nl/vrijwilliger-informatie. Accessed: 29/5/2023.

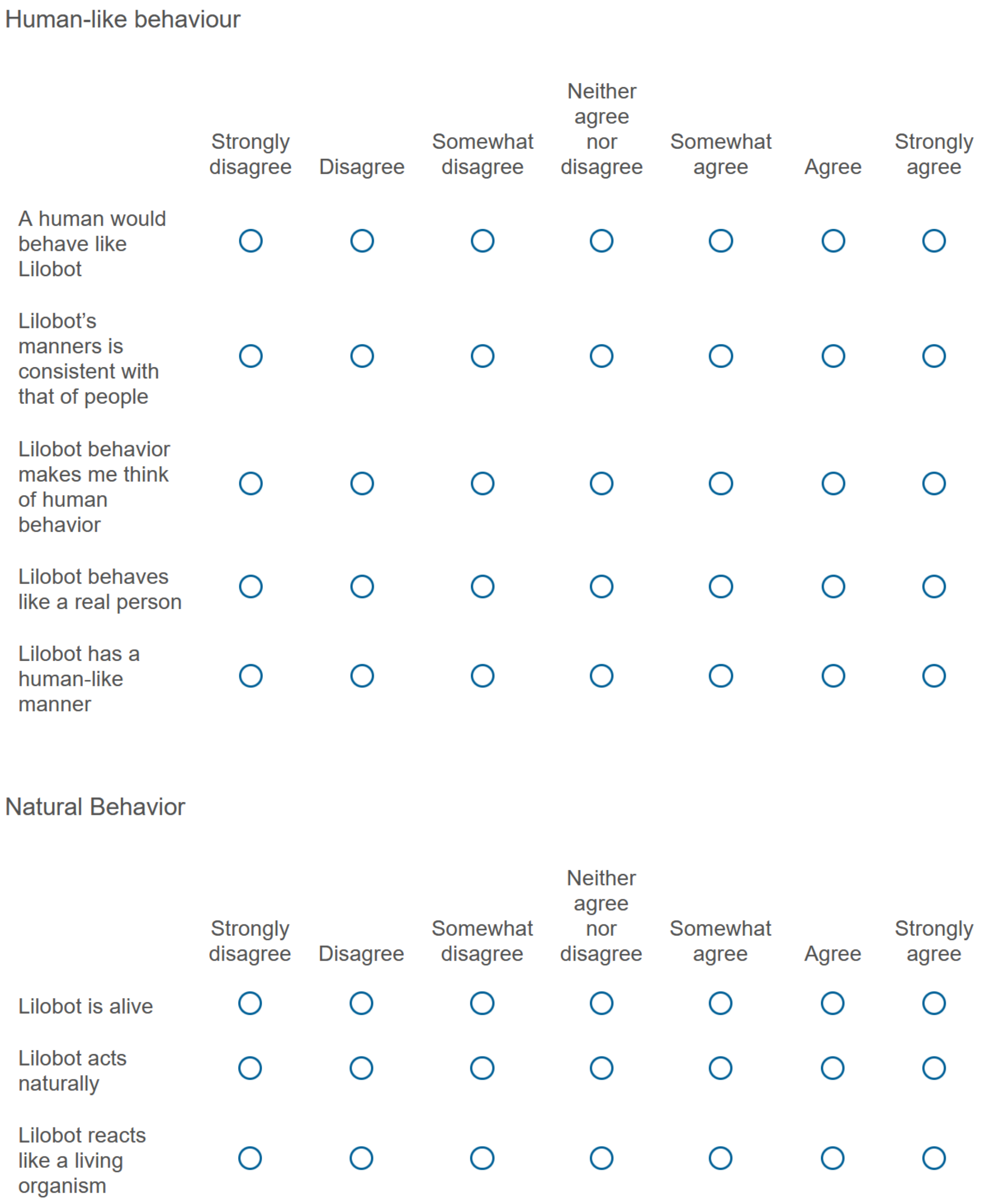
[15] Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. Leveraging large language models to power chatbots for collecting user self-reported data. (arXiv:2301.05843), Jan 2023. doi: 10.48550/arXiv.2301.05843. URL http://arxiv.org/abs/2301.05843. arXiv:2301.05843 [cs].

[16] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from language models, Dec 2021. URL https://arxiv.org/abs/2112.04359v1.

A Interview Questions

1. What are the main aspects of Lilobot's personality that influenced your believability?
2. Imagine that instead of Lilobot you were talking to a real child who was bullied, what are some similarities and differences you think you would notice?
3. How would your personal feelings and responses differ?
4. How do you think the child's responses would differ?
5. Is Lilobot's behaviour suitable for its role as a chatbot for training helpline workers?

B Believability Survey



Human-like behaviour

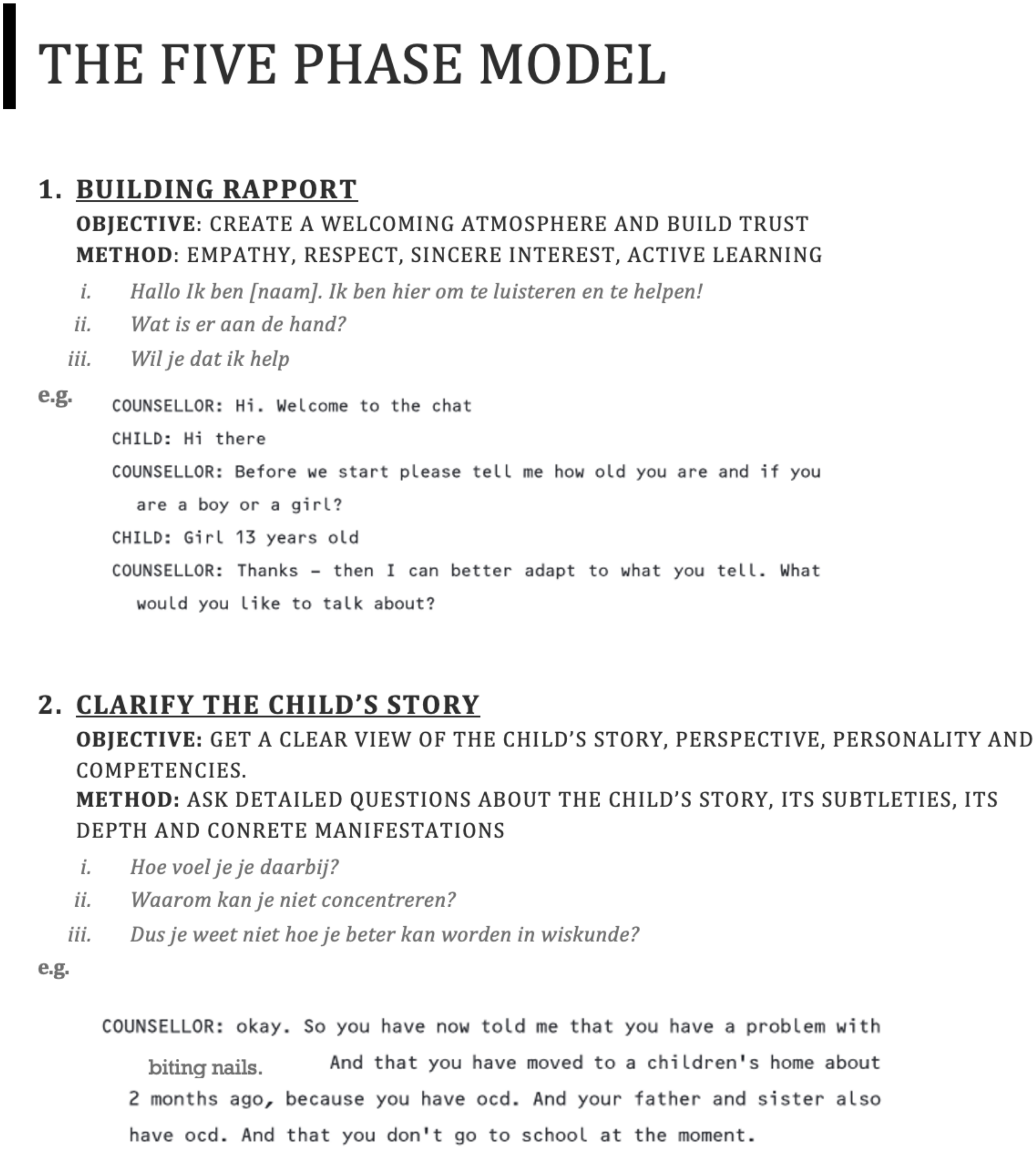
	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
A human would behave like Lilobot	○	○	○	○	○	○	○
Lilobot’s manners is consistent with that of people	○	○	○	○	○	○	○
Lilobot behavior makes me think of human behavior	○	○	○	○	○	○	○
Lilobot behaves like a real person	○	○	○	○	○	○	○
Lilobot has a human-like manner	○	○	○	○	○	○	○

Natural Behavior

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
Lilobot is alive	○	○	○	○	○	○	○
Lilobot acts naturally	○	○	○	○	○	○	○
Lilobot reacts like a living organism	○	○	○	○	○	○	○

Figure 3: The believability survey based on the artificial-social-agent questionnaire [8] that was given to the study participants.

C Example Conversation Transcript



THE FIVE PHASE MODEL

1. <u>BUILDING RAPPORT</u>

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. <u>CLARIFY THE CHILD'S STORY</u>

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