The effect of anthropomorphism in conversational agents.

Jesse Jansen, Sihang Qui, Ujwal Gadiraju, Jie Yang

TU Delft

Abstract

Conversational agents are being used more often. Many conversational agents have been designed to make them more humanlike to increase the engagement of the user. This research is looking into a couple of aspects of anthropomorphism to determine whether these aspects improve the engagement of the users. These aspects are typos with time delay and empathy. 3 conversational agents were created for this research, one control group and one for both aspects. For each conversational agent a group of volunteers tested the bots and filled in a survey after the conversation. While the conversational agent with typos and time delay didn't show a significant change, the conversational agent with empathy did. When testing the control conversational agent and the conversational agent with typos and time delay multiple participants told that they felt ignored when the conversational agent doesn't show empathy, this implies that not having empathy has negative effects on the conversation. Therefore giving the conversational agent empathy results a better engagement of the user.

1 Introduction

The use of conversational agents is getting more and more common[8]. A conversational agent is "a software which users interact with using natural language"[1]. A conversational agent can be designed in many different ways [2] impacting how the user interaction goes and creating different levels of anthropomorphism. Anthropomorphism can be defined as how humanlike a conversational agent is. The more anthropomorphic a conversational agent is, the more it will appear to a user talking with the conversational agent that the user is talking to an actual human. It is important to have a high level engagement of the user when the user is having a conversation with a conversational agent, as a low level of engagement can result in boredom and potential sloppy work. [9]

Previous research suggest that increasing the levels of anthropomorphism would result in a higher level of satisfaction until the point where the slight imperfections give a feeling of uncanniness[3]. In the field of robotics this theory is called the Theory of the Uncanny Valley[4]. This theory is suggested to also apply for the field of conversational agents[3]. Previous research looks mostly into the different levels of anthropomorphism, but a lot less is known about what specific aspects of anthropomorphism improve the satisfaction or cause a feeling of uncanniness.

The goal of this research is to determine which aspects of anthropomorphism have a positive effect on the interaction with the user. To do this, this research will be looking into the research question:

RQ: To what extent can a conversational agent with different levels of anthropomorphism improve the satisfaction and engagement of the user?

The hypothesis that this research going to test is that by increasing the level of anthropomorphism in a conversational agent the satisfaction and engagement of the user in the conversation will increase until a certain point where the conversational agent becomes too

humanlike. For this hypothesis this research will only look at the textual aspects of anthropomorphism and not visual aspect.

This research will also look into specific aspects of anthropomorphism and for this there are a couple of sub questions.

RQ1 What impact does adding typos and a time delay between messages to a conversational agent has in terms of satisfaction and engagement of the user?

RQ2 What impact does adding empathy to a conversational agent has in terms of satisfaction and engagement of the user?

For the first sub question this research will look into how the user responds if the user has a conversation with a conversational agent that has a time delay and sometimes makes a small typo. When the conversational agent makes a typo he fixes it in a new message. This conversational agent will be compared to the user having a conversation with a conversational agent in the way the user would normally expect to have a conversation with a conversational agent. The small typo and the small delay should make it more like the user is actually talking to an actual human instead of a bot. How this bot is designed will be explained in the methodology section.

For the second sub question nearly the same thing has to be done, except the second conversation agent will have empathy to some degree for more humanlike responses. In both sub questions this research will look into how the user responds and if this is an improvement. How this bot is designed will be explained in the methodology section.

The rationale behind RQ1 is that it will make the conversational agent type more humanlike, without changing what the conversational agent says. Previous research suggest that time delay has a positive effect[10], this research combines the time delay with typos. If this is an improvement in terms of engagement and satisfaction this can be used in the future to improve conversational agents without having to change the conversation. The rationale behind RQ2 is that empathy is an important factor in how normal conversations go. Having the ability to show empathy improves conversations when giving advice[5], if this also counts for regular conversations with conversational agents this could potentially improve future conversational agents.

After performing the research the results did not show significant improvements for both aspects. Multiple participants that tested a conversational agent without empathy complained that they felt ignored when the conversational agent did not show empathy. This implies that it is highly recommended to give a conversational agent the ability to show empathy.

The rest of this paper will have the following structure. Section 2 will explain the methodology of this research. Then Section 3 will show the actual research and the results. Section 4 will explain more about what the results imply looking at the research questions and what the limitations of this research are. Then section 5 will explain how this research dealt with ethics. And finally Section 6 will provide a conclusion and describe what future research can be done.

2 Methodology

The methodology that is used by this research is doing between-user-tests on different conversational agents. As there are 2 aspects of anthropomorphism tested, there is one group of users that need to do the test with a conversational agent that has one of these aspects. Besides these 2 groups a third group is needed, this group will have none of the 2 aspects that this research will look into, this group is called the control group. The control group is important to have to get valid results in the experimental groups[5]. How these conversational agents are

designed will be described in the next subsection. As the control group can be used for both research groups only 1 control group is needed, which will provide a base line for the research. Each user is asked to have a short conversation with one of the conversational agents and at the end of the conversation the conversational agent will ask the user to fill in a short survey. Which user gets which conversational agent is randomised. Finally the results of the survey are compared to see whether the conversations with the conversational agents that had an added aspect of anthropomorphism had an improvement or not. Each conversational agent should be tested with by 10 users to get a valid result.

2.1 The conversational agents used in this research

As mentioned before this research uses a total of 3 different conversational agents. Each conversational agent will have the same conversation with the user. The conversations have to be as similar as possible to prevent other aspects of the conversation impacting the results, therefore the only difference between the conversations should be the aspect of anthropomorphism that is being tested.

2.1.1 Control conversational agent

The first conversational agent is the control group. This conversational agent will not have any of the aspects that are being tested. As mentioned before this conversational agent is being tested by the user to provide a base line. This means that the results of the other conversational agents can be compared to this conversational agent to see if they scored better[6].

The conversation that the participant has is a short conversation where 2 students get to know each other. During this conversation the conversational agent asks simple questions, which a student can ask another student when they talk for the first time, for example “Which study do you follow”. These questions are asked in a set order and they are all the same for each different conversational agent. The beginning of each of the conversations can be seen in each corresponding figures. In the actual conversations the buttons with good/not good will be removed after the participant presses one of them, but for the purpose of showing these buttons they haven’t been removed when making the screenshots.

Figure 1: A conversation with the control bot

2.1.2 Conversational agent with typos and time delay

This conversational agent is the first that has an added aspect of anthropomorphism. This aspect that has been added in this conversational agent is

the addition of typos and time delay to simulate that the user is talking to a human. Each message this conversational agent sends will have a short time delay instead of instantly sending the message, this simulates the conversational agent needing time to type the message. The length of the delay varies based on the length of the message, as a longer message needs a little more time to type. Besides this delay a few messages of this conversational agent will have a typo added, not every message. This typo is a word that is missing a letter, an example of this can be seen in figure 2. After the conversational agent sends a message with the typo it will send another message fixing the typo. This conversational agent is used to answer RQ1.

Figure 2: A conversation with a bot with a typo

2.1.3 Conversational agent with empathy

Figure 3: A conversation with a bot with empathy

The third conversational agent is the conversational agent with empathy. This conversational agent will send some additional messages to show empathy. For some questions that the conversational agent asks the user has to pick one of the preselected answers and based on this answer the conversational agent will send an extra message to show empathy. A specific example of a question where this is used is the question “how are you feeling today” if the users selects “not good” then the conversational agent will answer with “I’m sorry to hear that” before moving on to the next question, this can be seen in the figure 3. The control conversational agent will simply ask the

next question without responding to the response of the user. This conversational agent is used to answer RQ2

2.2 The survey

At the end of the conversation the conversational agent ask the participant to fill in a survey. This survey asks the participant how much he can relate to certain statements about the conversation on a scale of 1 to 6, where 1 is strongly disagree and 6 is strongly agree. The results of this survey will be compared over the different conversational agents to see if there is an improvement in terms of engagement. To measure the engagement of the user the User Engagement Scale (UES)[7] is used. At the end of the survey there is an open question where the participant can leave any feedback.

3 Research and results

3.1 The research

After creating the chatbots it is time to test them with the group of volunteers. Before the volunteer starts the research he or she gets a short instruction. In this instruction the volunteer is told that he will have a conversation with a conversational agent and that there will be a survey at the end. The volunteer is also told that the data collected is completely anonymous and only the data of the survey is stored. Then the volunteer gets a randomly assigned conversation agent, either the control or one of the two with an additional aspect, and has the conversation with the conversational agent followed by the survey. In this survey the user is asked to score statements from 1 to 6 and an open question in the end:

Q1 I think the conversation feels natural.

Q2 The conversation was how I expect to have a conversation with a chatbot.

Q3 I liked the conversation.

Q4 The conversation felt like talking to an actual human.

Q5 Do you have any verbal feedback?

3.2 The results

For the control group the sample size was 8 participants. The result of this group can be found in table 1.

	1	2	3	4	5	6	Average
q1	0	5	3	0	0	0	2,38
q2	0	2	4	2	0	0	3
q3	0	2	4	2	0	0	3
q4	2	4	2	0	0	0	2

Table 1 The results of the control conversational agent

This group gave the following verbal feedback: “could use some response on what the user says.” And “I felt ignored when the bot didn’t respond to me saying I wasn’t feeling good.”

3.2.1 The results with the typos and delay

The group with typos and delay had a sample size of 8. The results of this group can be found table 2.

	1	2	3	4	5	6	Average
q1	0	6	2	0	0	0	2,25
q2	0	2	2	2	2	0	3,5
q3	0	3	1	4	0	0	3,13
q4	4	0	2	2	0	0	2,25

Table 2 The results of the conversational agent with typos and delay

This group gave the following verbal feedback: “Could use some response on what the user says.” “I felt ignored when the bot didn’t respond to me saying I wasn’t feeling good.” And “It was funny that there was a typo”

There is only a very small difference compared to the control bot with how the participant rated the bot. This difference can be explained by the fact that there was a very small sample size. This conversational agent did however score higher at the question if the conversation was how the user expects to have a conversation with a chatbot. This implies that this is more how a user would expect a chatbot to behave.

When looking at the verbal feedback it is worth noting that at the conversational agent with empathy one participant noted that the messages were quit fast after each other, which suggest that that participant would prefer a short time delay within messages, which was also suggested by previous research[10]. However this is only the verbal feedback of a single participant and therefore in this research it is not enough to make a valid claim that this is a significant improvement.

3.2.2 The results with the empathy

The group with empathy had a sample size of 10. The results of this group can be found in the table 3.

	1	2	3	4	5	6	Average
q1	0	6	4	0	0	0	2,4
q2	0	4	2	4	0	0	3
q3	0	2	4	2	2	0	3,4
q4	4	2	4	0	0	0	2

Table 3 The results of the conversational agent with empathy

This group gave the following verbal feedback: "The messages were a bit quick after each other"

There is also only a very small difference compared to the control bot. The difference can be explained by the very small sample size. The only improvement that is worth noting is that the participants gave a higher score at the question if they liked the conversation, however given the small sample size this improvement is not very significant.

When looking at the verbal feedback there is a clear difference, however, this is implied by what the participants had to say about the other conversational agents and not about this conversational agent. For both the control bot and the bot with typos and delay multiple participants responded that they felt ignored when the conversational agent did not respond to them pressing the not good button in the question "how are you feeling today?" For the conversational agent with the empathy none of the participants mentioned anything about this question. This implies that if a conversational agent is not able to show empathy, this impacts the engagement of the user in a negative way, therefore adding empathy has a positive effect.

4 Discussion

After collecting all the results it is important to determine what these results imply, why this is the case and what the limitations of this research were.

4.1 Implications

As mentioned in the previous section there is a clear indication that having a conversational agent that is not able to show empathy has a negative impact on the engagement of the user. This implies that it is highly recommended to add empathy to conversational agents.

Adding typos and time delay has an insignificant impact on the engagement of the user. A single participant did mention that messages were too fast after each other while testing one of the other conversational agent, however this would require more research to make this a valid claim.

Comparing these results to the hypothesis, the hypothesis was correct when applied to RQ2. Based on the verbal feedback of the participants the engagement and satisfaction were improved by adding empathy to the conversational agent. As the level of anthropomorphism was increased by adding an aspect of anthropomorphism to the conversational agent and the engagement and satisfaction was improved this confirms that the hypothesis was correct.

When comparing the hypothesis to RQ1 the hypothesis was not confirmed to be correct as the results did not show a significant improvement.

4.2 Limitations

The results of this study suggest that adding empathy will improve the

satisfaction and engagement of users talking with a conversational agent. For a fair comparison the conversations were tried to be nearly exactly the same. However the conversational agents tried to read what the participants say, to be able to respond to them, by using regular expressions. If the conversational agent couldn't read the user input it would send a response with "I'm sorry I can't read that". As the conversations were not stored for privacy reasons it has not been recorded if some participants got to see this message a lot more than others and therefore the impact of seeing this message can't be determined.

The fact that the sample size was quit small, unfortunately smaller than originally intended, could also have an impact on the results. In a small sample size the more extreme results, which are participants that gave a much higher or lower score in the survey, have a much larger impact on the average result then when the sample size is much larger.

The final limitation is the design of the conversation. The conversation is aimed at students and most participants were technical students with a background in computer science. The conversation was designed to be informal, for adding typos the result would most likely be different in a formal conversation. An attempt was made to design the conversation in a way that both empathy and typos would not greatly impact the conversation, but it was not verified whether the research used the right design.

5 Responsible research

For every research that uses human participants it is important to discuss the ethics in the research. For this research the users will first have a conversation with the conversational agent and then fill in a survey. During the conversation there will not be any data saved, all messages the user sends to the conversational agent will be deleted immediately after the conversation. The data of the survey is stored as this is the actual research data. The survey is completely anonymous, the only data about the user that is stored in the survey is which version of the conversational agent the user has had a conversation with. The user doesn't know which version he got. Before the research starts the participant will be told which data is stored and which isn't. The participant is given the option to end the participation at any point, if the user decides to end the participation before finishing the task there will not be any data stored. As the results of the survey will be anonymous it will not be possible to delete the data after the participation has finished.

To make this research reproducible the survey is included. The conversational agent used in this research were created using the python telegram framework.

The survey had the following questions:

Q1 I think the conversation feels natural.

Q2 The conversation was how I expect to have a conversation with a chatbot.

Q3 I liked the conversation.

Q4 The conversation felt like talking to an actual human.

Q5 Do you have any verbal feedback.

6 Conclusion

6.1 Conclusion

The goal of this research was to determine if adding different aspects of anthropomorphism to conversational agents would improve the engagement of the users. In order to do this 3 conversational agents were tested. A control conversational agent, a conversational agent with a time delay between messages and typos and finally a conversational agent that was able to show empathy to some degree. While there was no significant difference

between the control and the conversational agent with time delay and typos the participants noted that they felt ignored if the conversational agent does not show empathy, which implies that adding empathy improves the engagement of the users.

6.3 Future research

In this field there is still quit a lot of possible research to be done. If this research were to be continued a larger sample size and possible a longer conversation would be a good start. The next step would be combining the aspects that this research looked into. This would result in a conversational agent with both typos, time delay and the ability to show empathy.

The next step is looking into other aspects of textual anthropomorphism. Possible aspect to research are humour and emotions, but there are many more possible aspects to research. These aspects can also be combined in more future research.

Another possible future research is looking into different levels of anthropomorphism in a single aspect.

References

[1] McTear M, Callejas Z, Griol D (2016) The conversational interface: talking to smart devices. Springer, Basel

[2] Amir Shevat. Designing bots: Creating conversational experiences. O'Reilly Media, 2017.

[3] Diederich S, Brendel AB, Kolbe LM (2020) Designing anthropomorphic Enterprise Conversational Agents, Göttingen

[4] Mori M (1970) The Uncanny Valley. Energy

[5] Liu B, Sundar SS (2018) Should Machines Express Sympathy and Empathy? Experiments with a Health Advice Chatbot.

[6] Bruchmann K (2017) Compared to What? The Importance of Control Groups in Social Comparison Research.

[7] O'Brien HL, Cairns P, Hall M (2018) A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form

[8] Yang X, Aurisicchio M, Baxter W (2019) Understanding Affective Experiences With Conversational Agents.

[9] Qiu S, Gadiraju U, Bozzon A (2020) Improving Worker Engagement Through Conversational Microtask Crowdsourcing.

[10] Gnewuch U, Morana S, Adam MTP, Maedche A (2018) Faster Is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction, Portsmouth