Putting Chatbot Interactions on the Right TrAKX: Examining Whether Slight Delays in Conversation Increase a User's Trust in Recommendation Systems

- Khellaf, Tarek (t.khellaf@students.uu.nl)

- Daggelinckx, Anneline (a.daggelinckx@students.uu.nl)

- Zhang, Xinyu (x.zhang16@students.uu.nl)

- Schoneveld, G, Kuil (k.g.schoneveld@students.uu.nl)

# Research question and overall method

Will a short delay in a chatbot's responses make the user trust the chatbot and it’s decisions more, thereby leading to an increase in the participants' overall satisfaction with their interaction? More specifically, do slight delays in text conversations increase the trust a user places in their interlocutor? To explore this question, we posted our TrAKX chatbot online, designed a series of preference lists and two questionnaires that will be provided to our participants, and prepared an appropriate data analysis model. We plan to invite 30-40 participants and send them a link to our test which integrates the chatbot and questionnaires with the necessary hints. For each test, we will provide 2 pairs of preference lists to the participant; the notable difference is that one pair can be directly matched in the database while the other needs an alternative choice. The participant will chat with botA to find a restaurant according to preference lists 1a and 2a, and with botB according to preference lists 1b and 2b. The order of botA and botB is randomly arranged, only one of which has immediate responses. Participants then provide information about the target restaurant (e.g. address, phone number) to make sure they perform the request we expected. After chatting with each bot, the participant will be asked to fill in a questionnaire, logging the participant’s age, their trust with each bot, and extra notable feelings. We ask multiple questions related to trustworthiness to increase our confidence in the relevant answers and encourage extreme values to be averaged by other answers. This should clarify any overall trends we find.

# Data collection

We haven’t collected data yet but focussed on setting up the entire experiment. We plan to start collecting data by the end of this week. Since we have set up the experiment entirely online, we only have to send a link to the participants, and they will be able to go easily from page to page, testing the bot’s and filling in the forms. We think that it won’t be too hard to find participants this way, because we don’t have to interfere and they can do the experiment whenever they have time.

# Data analysis

We will collect our data via a google form. This form exports the data to an excel file. In this excel file, we can take the average of chosen numbers of the different analogous questions in the form. Since we use a separate form for both bot settings, we will have two such excel files. Because we want to do a paired test, it is important to pair the results of the two files. For this reason, we will give users a user-id that will be filled into the form automatically when they enter our test website. For the data analysis, we have made a script in R. This script reads in the excel files and makes boxplots out of the averaged question results. To check whether there is a significant difference, we do a one-sided, paired t-test. The test is paired because we have a within-subject design, and is one-sided because our null hypothesis states that there is no significant difference, and our alternative hypothesis states that people will find the bot with the delayed response more trustworthy. To end with, and as an extra, we plot a histogram of the participant responses where they directly have to choose between the first and second bot. We tested this script with made-up data.

# Distribution of labor

This week, we spit up the work a bit. Kuil focussed on the questionnaires, Tarek set up the online experiment, Xinyu searched for suitable user preferences that the user will have to ask to the bot, and Anneline set up the preprocessing and data analysis of the experiment. Of course, we discussed every aspect together during calls.

# Questions

* If we find some participants didn’t get the right restaurant information, should we just discard their answers? Or should we do some special processing with these data？
* Besides age, are there any other personal characters you think we might pay attention to?