**TrAKX (our chatbot): User Study**

- 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**

Will a short delay in the system's responses make the conversation feel more human-like, thereby leading to an increase in the participants' overall satisfaction with their interaction? The number of user-chatbot interactions have increased dramatically as businesses include such systems on their websites. We therefore believe that this question is worth exploring. Improving customer satisfaction could potentially help improve user acceptance of- and trust toward chatbots. The paper *Faster Is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction*, found that “dynamic response delays can not only increase users’ perception of humanness and social presence, but also lead to greater satisfaction with the overall chatbot interaction”. We will therefore attempt to replicate the study’s results and test whether or not the results are applicable to our chatbots’ use case.

**Method**  
Participants  
For our experiment, we will ideally invite 30 to 40 participants to populate a sample strong enough to form trends in the results. We will manage the experiment online, including online chatbot conversation and questionnaire answering. Our aim is to invite a wide range of people of different ages and backgrounds. The ages of participants would be particularly useful for analysis if our experiment didn't get expected results in all ages/backgrounds range. A requirement to participate in our experiment is that you are familiar with the English language since TrAKX requires English communication. In the case of an online experiment, we have begun recruiting participants via our social media and acquaintances.

Materials  
We have posted our bot to an online platform so participants can click a link and join the experiment. Since the task is straightforward, we won’t give the user training time. We will present the user with an explanation of what kind of restaurant the user has to ask for. Thereafter, the user can chat with TrAKX to get a recommendation. For each setting, the user will have two lists of preferences to give TrAKX, one short and one long. This is because TrAKX often provides the same utterances and with too many conversations, the user might get bored; potentially biasing our results. Preference lists we provide in different versions of bots would be similar (e.g. both can find a suitable recommendation or both need alternative choice) so that preferences wouldn’t bias our results. After the conversations, the users will be provided with a questionnaire.

Design & Procedure  
Due to the time and resource constraints, however, we will not be able to perform a large test upon two separate groups for comparison. Given the relatively low number of participants, testing the subjects’ impressions when given both versions (one group with delay first, the other without delay first) will result in a larger number of data points and be more informative. By altering the complexity of the preferences, we can manipulate the difficulty of the search and ensure that the potential error between short- and long preference lists will remain null.

We will divide the participants into two groups, each starting with one version of the bot followed by a questionnaire and then vice versa. A conversation with TrAKX is never very long, so the experiment itself will take about 5-10 minutes. We estimate that filling in the questionnaires will take another 5 minutes, which brings us to 10-15 minutes in total.

Measurements  
We plan to have the participants converse with the system and then rate how they felt during the interaction. From this information, we will be able to extrapolate the statistically significant differences between delayed and immediate responses with respect to the perceived trustworthiness of each.We have considered including other measurable factors, such as the length of conversations or the average number of words per user utterance and how they differ between the two versions.

**Hypothesis**  
We hypothesize that participants will find TrAKX more trustworthy if there is a time delay before the system utterance, while the lack of delay might give the users the feeling of being rushed. This aligns with the hypothesis found in *Faster is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction.*

**progress**

(One paragraph on your progress in terms of data collection: how many participants have you collected data on? How many more do you plan to do? Is this realistic within the limited scope? )

**data analysis**

(Have you already made progress on pre-processing the data such that you can quickly do the main analysis next week? If you have not done that yet, what are you planning to do?)

**Work Distribution**We aim to distribute the labor evenly between our group members. We will discuss and execute the experiment together and everyone will try to find an equal amount of participants. Of course, not everyone will do the statistical calculations, but we will discuss them together. Just as we did in Part 1, we will write the report first with bullet points, so everyone can have a say in every part, and then integrate these points into a coherent text.

Question

Do you think it is necessary to check whether the participants have got the right information? If we find some participants didn’t get the right information, should we just discard their answers？ Or do some special processing with these data？

About participants’ gender?

**Reference**

Gnewuch, Ulrich & Morana, Stefan & Adam, Marc & Maedche, Alexander. (2018). Faster Is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction.