The Effect of Delay-Handling Strategies on Perceived Usability of a Chatbot

Lisa Aaldijk, Gijs van Nieuwkoop, Marc Overbeek and Olivier Vroom
Group 6
8/11/2024

1 Introduction

Customers facing an issue do not just want their issue solved as quickly as possible, they also want to have a good experience while interacting with a customer service agent. Chatbots and dialog systems are becoming increasingly used to provide faster and more human-like responses and solutions. For simplicity, the term *chatbot* will be used to refer to both chatbots and dialog systems throughout this paper, as done by others in the field (Caldarini et al., 2022).

Before the launch of ChatGPT, the adoption of chatbots was slower than anticipated. This was largely due to their perceived lack of human-like behavior, resulting in robotic interactions (Grudin & Jacques, 2019; Ryan M. Schuetzler & Giboney, 2020). To address this, designers began incorporating "social cues" based on the "Computers Are Social Actors" (CASA) paradigm. It suggests that humans apply similar social heuristics to computers as they do to other humans (Feine et al., 2019; Nass & Moon, 2000). For example, simple human-like behaviors, such as a typing indicator, can induce patience in users. This enhances perceived usability by fostering a sense of social presence. Research on such cues, including assigning names or personalities to chatbots, has shown positive impacts. However, studies specifically on delay handling remain limited (Gnewuch et al., 2018). Even state-of-the-art large language models (LLMs) require time to generate responses, and while typing indicators have been shown to increase perceived social presence, these effects were mainly seen in users new to chatbots, who may not be aware that a chatbot is not actually typing like a human would (Gnewuch et al., 2018). Experienced users, by contrast, found purposeful delays frustrating since they expected faster responses (Gnewuch et al., 2022).

With ChatGPT's launch in November 2022 and subsequent releases of similar LLM-based chatbots, chatbots gained enhanced human-likeness (Caldarini et al., 2022). However, increased complexity has led to longer response times. This introduces new questions about optimal delay handling—should users see a blank screen, a typing indicator, or the response appearing word by word? While some research on delay handling exists, much of it predates the widespread adoption of LLM-based chatbots, whose advanced capabilities have reshaped societal perceptions of chatbots.

This led to the following research question: *What is the effect of different forms of delay handling on the perceived usability of a chatbot?*

The following hypotheses were formulated:

Hypothesis 1: *Delay handling "typing", in which chatbot responses are outputted character by charachter will result in the highest perceived usability.*

Hypothesis 2: *Delay handling by writing dots during delay will yield a higher perceived usability compared to not handling delay at all.*

These hypotheses are based on the CASA paradigm, which suggests that incorporating social cues, such as typing, can increase patience and potentially enhance perceived usability (Feine et al., 2019; Nass & Moon, 2000). Additionally, another intuitive explanation for these hypotheses is that in both cases, users are exposed to some stimuli whilst waiting for the chatbot's response,

potentially decreasing frustrations about the waiting time.

Research has shown that adding a typing indicator can help increase the social presence of chatbots (which ultimately improves usability), but that this only occurs for users with little to no prior experience with chatbots (Gnewuch et al., 2018). However, this research was done in 2018, before the launch and widespread adoption of ChatGPT. It is interesting to explore whether this relationship still holds, now that the use of chatbots is more common. This leads to the following sub-question: *How does a user's prior experience with chatbots influence the effect of different delay handling strategies on their perception of usability?*

By investigating and answering these research questions, companies can make more informed decisions on what type of delay handling to implement for their chatbots in order to maximize perceived usability, and ultimately user experience.

2 Method

2.1 Participants

The experiment was conducted with forty participants, selected using convenience sampling. All participants were proficient in English to make sure they could understand and interact with the chatbot. There was also some degree of diversity in the participants. The participants were of different ages, educational backgrounds, and genders, and had different levels of prior experience with chatbots.

To maintain ethical standards, the ethical quick scan survey was filled out and informed consent was obtained from the participants before the start of the experiment.

2.2 Experimental design

The experiment was a within-subject comparison between three conditions. In the first condition, the chatbot did not generate any output during the delay period. This condition will also be referred to as the "baseline" condition further on. In the second condition, the chatbot was outputting three dots one by one during the delay, with constant time intervals between each dot. In the third and final condition, the chatbot generated its response character by character to fill up the time delay. The total delay in all conditions was 2 seconds.

The experiment had three different restaurant search tasks. When done most efficiently, the completion of all three of these tasks required the exact same amount of user utterances.

Each participant encountered all chatbot conditions and all restaurant search tasks, but the order for both was randomized to counterbalance learning effects.

Then, after each interaction with the chatbot, the perceived usability under the specific condition was measured. After having completed all interac-

tions, participants filled out a demographical information questionnaire indicating their level of previous experience using chatbots and their age. Finally, the perceived usability was compared across the conditions.

2.3 Materials

In order to perform the experiment, a mixed-initiative chatbot was developed. The chatbot was capable of recommending certain restaurants from a database, after taking in the user's preferences. When using the chatbot, a user answers questions asked by the chatbot in order to receive a restaurant suggestion. If desired, a user could request more information about the suggested restaurant. When the user was satisfied, typing "bye" concluded the interaction.

2.4 Measurements

Measuring perceived usability is a difficult task since it is a hard concept to test and quantify. In this research, the questionnaire by Holmes et al. (2019) was used because this questionnaire is proven to test for the perceived usability of chatbots. It consists of 16 questions evaluating the perceived usability of the chatbot. With the responses, a perceived usability score between 0 and 100 can be calculated using the formula from Holmes et al. (2019). An outcome of 0 in the questionnaire means dissatisfied and 100 means fully satisfied. These scores could be compared across conditions and were therefore useful for this research.

From each participant, a perceived usability score for each condition, in addition to their age and prior experience with chatbots was obtained. To determine whether there was a significant difference between the conditions' usability scores, a repeated measures ANOVA was done. This statistical test was deemed appropriate for the performed within-subject experiment with three conditions.

The ANOVA was done using the chatbot condition as the independent variable and the calculated perceived usability score as the dependent variable to answer the central research question of whether there is a significant difference in perceived usability between different forms of delay handling.

To answer the subquestion, it was tested whether there is a significant effect of prior experience with chatbots on the perceived usability for each condition. To this end, for the usability scores obtained in each of the three conditions, a Pearson correlation test was done between prior experience and perceived usability.

2.5 Procedure

The procedure for each participant task can be seen in Figure 1. First, participants were asked for consent and it was explained to the participants that there would be three rounds where they were asked to find a specific kind of restaurant. Which restaurant to find was written in the task they received right before each round.

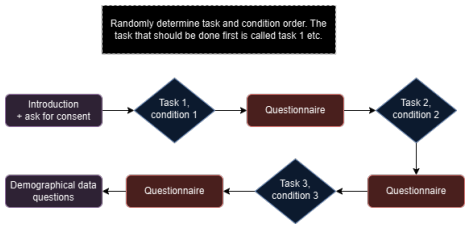

Figure 1: Participant task procedure

Second, the participant received a task and after the participant was done using the chatbot they filled out the sixteen questions of the questionnaire by Holmes et al. (2019).

The second step repeated itself twice with the chatbot in a different condition so that each participant interacted with the chatbot with every form of delay handling.

Last, the participant filled out a demographical information questionnaire.

3 Results

The results of the experiment are shown in Figure 2. Here we plotted the usability scores for the three different types of delay handling. The results of the repeated measures ANOVA showed that the baseline ($M = 62.85$, $SD = 17.40$), dots ($M = 61.23$, $SD = 16.06$), and typing ($M = 67.40$, $SD = 15.34$) conditions were not significant ($p = 0.1482$, $F = 1.9689$) for the perceived usability. Secondly, Pearson's correlation tests were done. The results of these tests showed that under the baseline ($p = 0.14$), typing ($p = 0.96$), and dots ($p = 0.85$) conditions, no significant effect was observed between the participants' level of experience with chatbots and the perceived usability scores that were given.

4 Discussion

This study aimed to examine the effect of different forms of delay handling on the perceived usability of chatbots. Furthermore, the role that previous experience with chatbots had on this effect was also investigated. These questions were addressed by having participants interact with a chatbot, varying between three different ways that the chatbot handled a delay in its responses, and subsequently assessing perceived usability. No significant effect was observed for the different delay handling types on the perceived usability of the chatbot, whilst participants' previous experience with chatbots in turn also did not have a significant effect on this relation.

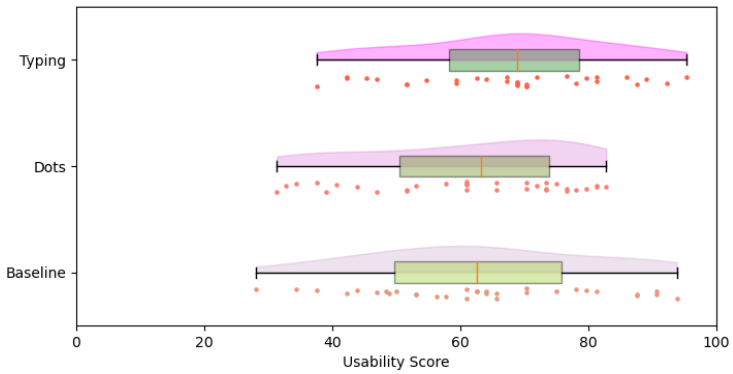

Figure 2: Raincloud plot showing the usability scores, their corresponding boxplots, and their probability density distributions across the three different ways in which delay was handled during experiments.

4.1 Implications

The outcomes of this study provide insights into how different delay-handling strategies influence the perceived usability of chatbots. More specifically, the findings suggest that there was no significant difference in perceived usability across the three different ways in which delay was handled in the experimental setup. This contrasts with previous research that associates human-like behaviors with increased usability (Gnewuch et al., 2018), since generating responses character by character resembles the way in which humans would type out responses more closely compared to outputting a full response at once.

More generally, this implication undermines the CASA paradigm that was mentioned in the introduction, which posits that people interact with computers as they would with other humans, attributing human-like qualities to them and applying similar social cues and expectations (Feine et al., 2019; Nass & Moon, 2000). The fact that this study found no significant difference in perceived usability across the different delay-handling strategies suggests that social cues might not actually play a role in conversations with chatbots.

Additionally, the fact that no significant correlations were found between participants' previous experience with chatbots and their assigned usability scores, contrasts with the findings of Gnewuch et al. (2018), that suggest that users with no prior experience with chatbots experience a higher level of usability when a typing indicator is used. This could be a sign that the widespread adoption of LLM-based chatbots since then has reshaped the way that people perceive and rate chatbots.

In practical terms, insights from this study are relevant for parties that utilize chatbots in their operational flow. Since literature has shown that the existence of delays in chatbot responses leads to a less enjoyable user experience

(Gnewuch et al., 2022), dealing with this delay in a conscious and well-informed way is important. However, based on the findings of this study, the different strategies that were evaluated did not result in a difference in perceived usability.

4.2 Limitations

This study faced several limitations due to constraints in time and resources. First, and perhaps most importantly, the interactions between participants and the chatbot were likely too simplistic. The chatbot that was used had limited functionality, only which allowed for short and straightforward tasks during the experimental setup. This led to brief interactions, which likely restricted participants' ability to form the substantiated opinions necessary for accurately assessing the chatbot's usability.

Secondly, a larger total delay in the responses of the chatbot might have allowed potential differences across the delay handling conditions to manifest more clearly. When the total delay is larger, its effect on perceived usability is magnified (Gnewuch et al., 2022). Therefore, it seems plausible that the potential effects of the way that this delay is handled would also become more clear.

Finally, the reliance on a convenience sampling method likely limits the generalizability of the findings (Peterson & Merunka, 2014). Participants were not randomly selected and given their social proximity to authors, may have been exposed to chatbots and similar technologies on a level that is not reflective of the general population, reducing external validity. Random sampling in future studies would likely help draw conclusions on chatbot usability that could prove to be more robust and generally applicable.

While other limitations may also have influenced this study, these three were identified as particularly important. Further research into the role of delay handling in chatbot usability would likely benefit from addressing these limitations.

4.3 Conclusion

In contrast with the stated hypotheses, the outcomes of this study suggest that the different ways in which delays are handled by chatbots have no significant effect on their perceived usability. This suggests that the perceived usability of a chatbot would remain the same, regardless of whether the developers of the chatbot choose to output responses character by character, output dots during waiting times, or not output anything at all during a delay in the chatbot's response.

Additionally, the results of this study did not show any effect of people's previous experience with chatbots on whether different delay handling strategies influenced the perceived usability of a chatbot.

References

Caldarini, G., Jaf, S., & McGarry, K. (2022). A literature survey of recent advances in chatbots. *Information*, *13*(1). Retrieved from https://www.mdpi.com/2078-2489/13/1/41 doi: 10.3390/info13010041

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2019). A taxonomy of social cues for conversational agents. *International Journal of Human-Computer Studies*, *132*, 138-161. Retrieved from https://www.sciencedirect.com/science/article/pii/S1071581918305238 doi: https://doi.org/10.1016/j.ijhcs.2019.07.009

Gnewuch, U., Morana, S., Adam, M., & Maedche, A. (2018, 12). "the chatbot is typing ..." – the role of typing indicators in human-chatbot interaction.

Gnewuch, U., Morana, S., Adam, M. T. P., & Maedche, A. (2022, Dec 01). Opposing effects of response time in human–chatbot interaction. *Business & Information Systems Engineering*, *64*(6), 773-791. Retrieved from https://doi.org/10.1007/s12599-022-00755-x doi: 10.1007/s12599-022-00755-x

Grudin, J., & Jacques, R. (2019). Chatbots, humbots, and the quest for artificial general intelligence. In *Proceedings of the 2019 chi conference on human factors in computing systems* (p. 1–11). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3290605.3300439 doi: 10.1145/3290605.3300439

Holmes, S., Moorhead, A., Bond, R., Zheng, H., Coates, V., & Mctear, M. (2019). Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? In *Proceedings of the 31st european conference on cognitive ergonomics* (p. 207–214). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3335082.3335094 doi: 10.1145/3335082.3335094

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81-103. Retrieved from https://spssi.onlinelibrary.wiley.com/doi/abs/10.1111/0022-4537.00153 doi: https://doi.org/10.1111/0022-4537.00153

Peterson, R. A., & Merunka, D. R. (2014). Convenience samples of college students and research reproducibility. *Journal of Business Research*, *67*(5), 1035–1041.

Ryan M. Schuetzler, G. M. G., & Giboney, J. S. (2020). The impact of chatbot conversational skill on engagement and perceived humanness. *Journal of Management Information Systems*, *37*(3), 875–900. Retrieved from https://doi.org/10.1080/07421222.2020.1790204 doi: 10.1080/07421222.2020.1790204

5 Appendix: Contributions

Task	**Gijs**	**Lisa**	**Marc**	**Olivier**
Experimental Design and Research Questions	4 hours	4 hours	4 hours	2 hours
Literature and Academic Background	1 hours	0 hours	0 hours	4 hours
Result Collection	3 hours	4 hours	5 hours	6 hours
Data Processing and Analysis	3 hours	5.5 hours	3 hours	0 hours
Result Interpretation and Discussion	6 hours	0 hours	4 hours	1 hour
General Refinement of Final Report	2 hour	5.5 hours	1 hours	3 hours
Total	19 hours	19 hours	17 hours	16 hours