Disclosure in User Interactions with Mental Health Chatbots: Examining the Role of Self-Disclosure Techniques

Author: Yushan Shan
Supervisors: Ujwal Gadiraju, Esra de Groot
Delft University of Technology

CSE3000 Research Project

Background

Mental health issues are rising, yet access to professional care is limited. mHealth tools and chatbots aim to provide scalable, always-available support for emotional expression.

Previous work, including Liang et al. (2024) [1] , shows that chatbot self-disclosure can build trust and increase user openness. However, if not designed carefully, it may reduce comfort or harm credibility.

The specific effects of different self-disclosure styles on users' willingness to share sensitive mental health information remain underexplored.

Research Question

How do different levels of chatbot self-disclosure affect users' willingness to disclose personal information in a mental health support context?

Sub-questions:

- How does self-disclosure influence willingness across SDI dimensions?
- What effect does it have on user trust, comfort, and acceptance?

Measurement & Hypothesis

Measured Variables:

- *SDI*: Willingness to disclose (5 topics, 5-point Likert scale). Average score used.
- *ASAQ* (selected items): Trust, comfort, and perceived appropriateness.
- *Pre/Post Perceived Disclosure Willingness*: Change in general willingness to disclose health information.

Hypotheses:

- *H1a*: Emotional self-disclosure by the chatbot will lead to greater user willingness to disclose personal information than factual or no self-disclosure.
- *H1b*: Emotional self-disclosure will result in more positive user perceptions of the chatbot, including trust, comfort, coherence, and overall attitude toward the interaction.
- *H2*: Factual self-disclosure will lead to moderate improvements in user willingness to disclose and user perceptions of the chatbot, compared to the baseline (no self-disclosure).

Experiment Procedure

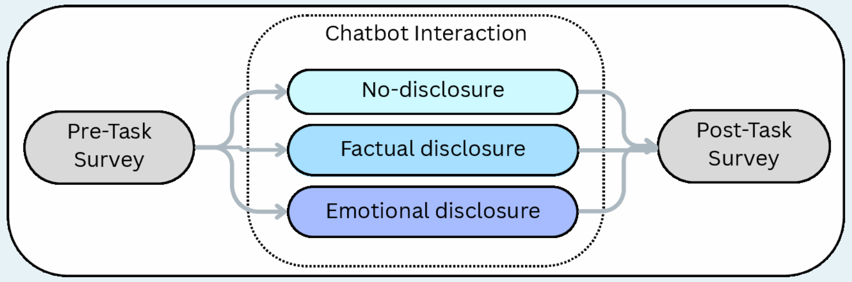
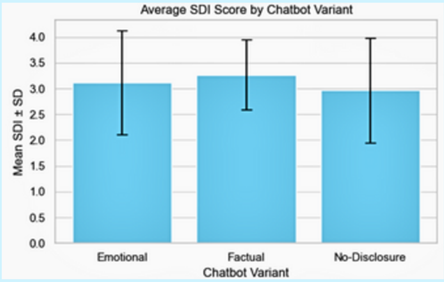


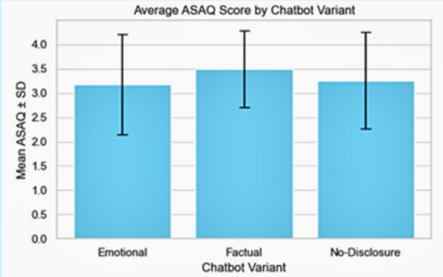
Figure 1. Chatbot Disclosure Experiment Design

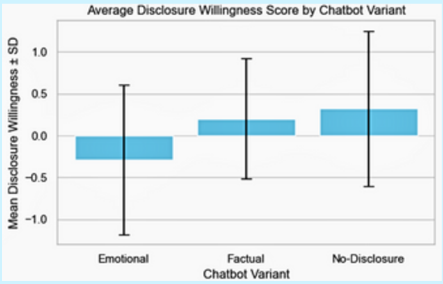
Participants first completed a pre-task survey covering demographics and their general willingness to disclose health information. They were then randomly assigned to interact with one of three chatbot variants: No Disclosure, Factual Disclosure, or Emotional Disclosure.

After the conversation, a post-task survey measured trust, comfort, and perceived appropriateness (ASAQ), as well as any changes in willingness to disclose. Participants also identified the chatbot's disclosure style to verify the experimental condition.

Result







Overall, 60% of participants correctly identified the chatbot's disclosure style. Recognition was highest for Emotional (71%) and No-Disclosure (65%), but lower for Factual (44%), indicating the manipulation was partially successful.

The Factual chatbot received the highest average scores for both SDI and ASAQ. The Emotional chatbot scored lowest on ASAQ and matched the No-Disclosure chatbot on SDI. A one-way ANOVA showed no significant differences across chatbot variants on these measures.

However, a significant difference was found in the pre/post change in willingness to disclose ($p = .013$). Post-hoc t-tests showed that the Emotional chatbot led to significantly lower willingness than both the Factual ($t = -2.42$, $p = .019$) and No-Disclosure ($t = -2.65$, $p = .010$) chatbots.

Discussion

The findings partially support the hypotheses. The Factual chatbot increased willingness to disclose and received the highest trust and comfort scores, supporting *H2*. However, the Emotional chatbot did not improve disclosure or perception, contradicting *H1a* and *H1b*. This contrasts with Liang et al. (2024) [1], who found emotional self-disclosure effective in movie and COVID-19 contexts.

These results underscore the importance of context. In mental health settings, emotional messages may feel intrusive or inauthentic during brief interactions. Without a sense of relationship, emotional content can reduce comfort. Factual or neutral disclosures may be more suitable in short-term conversations.

Limitations:

- *Resource Constraints:* Limited development time affected the chatbot quality, especially the Factual version, which was recognized correctly by only 44% of participants.
- *Short-Term Interaction*: Trust and comfort may take longer to develop; one-off sessions may underestimate emotional chatbot potential.
- *Role-Play Scenario:* Used to protect participants, but may have limited emotional realism and reduced authenticity of responses.

Future Work

Future studies should include a pilot phase to assess whether the chatbot's disclosure styles are clear and effective. Using LLM could help remove the need for a role-play scenario by enabling more natural, context-aware conversations. Collecting free-text responses may offer a more realistic view of actual disclosure behavior, although this would make it more difficult to apply structured measures like the SDI.

It would also be useful to explore adaptive disclosure strategies, where the chatbot gradually increases emotional depth as trust develops. Longitudinal studies could help examine how trust and comfort evolve over repeated interactions.

Reference

1. Liang, K.-H., Shi, W., Oh, Y. J., Wang, H.-C., Zhang, J., & Yu, Z. (2024). Dialoging Resonance in Human-Chatbot Conversation: How Users Perceive and Reciprocate Recommendation Chatbot's Self-Disclosure Strategy. Proceedings of the ACM on Human-Computer Interaction, 8(CSCW1), 1–28. https://doi.org/10.1145/3653691

Contact: y.shan-4@student.tudelft.nl ,Faculty of Computer Science, TU Delft