Do Privacy Policies Matter? Investigating Self-Disclosure in Mental Health Chatbots

A User Study on the Importance of Privacy and Question Sensitivity in Mental Health Chatbots

Manu Gautam[1]

Supervisor(s): Ujwal Gadiraju[1]**, Esra de Groot**[1]

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Manu Gautam
Final project course: CSE3000 Research Project
Thesis committee: Ujwal Gadiraju, Esra de Groot, Myrthe L Tielman

Abstract

Mental health chatbots are increasingly adopted to address shortage mental health services, by offering non-judgmental, always-available support. User self-disclosure is a critical factor which allows mental health chatbots to better understand users and provide more therapeutic experiences. Although prior work has explored how factors such as chatbot modality and tone affect self-disclosure, the role of privacy policies and how question sensitivity affects disclosure remains under examined. In this study, we investigate how privacy policies and the sensitivity of questions in voice-based mental health chatbots impacts user self-disclosure. Through a controlled user study, we explore whether the presence of a privacy policy leads to increased self-disclosure, whether question sensitivity influences self-disclosure willingness and whether there is any interaction effect between these two factors. Preliminary findings indicate that while providing a privacy policy did not significantly impact users' privacy understanding or willingness to self-disclose, question sensitivity notably influenced disclosure. Specifically, participants were more willing to disclose to low and medium sensitivity questions compared to high sensitivity. No interaction effect between the privacy policy and the question sensitivity was observed. Future research should expand participant pools, investigate self-disclosure in free-form interactions, and explore alternative methods of communicating privacy information for deeper insights into user perceptions regarding privacy, sensitivity and disclosure.

1 Introduction

Mental health disorders affect approximately 29.2% of people at least once in their lifetime, making them one of the most prevalent diseases in the world [33]. Despite this, mental health services remain significantly understaffed, especially in less economically developed countries. According to reports from the World Health Organization, up to 55% of people in economically developed countries and 85% of people in less economically developed countries do not have access or receive the mental health they need [4].

Mental health chatbots have emerged as a promising tool to help bridge this gap in mental health care accessibility. These chatbots offer 24-hour availability and convenience for their users, efficiently addressing the lack of accessibility and availability of mental health care. Furthermore, their perceived non-judgmental nature encourages users otherwise reluctant to seek other sources of help due to stigmatization to seek help [1].

Self-disclosure, the process by which a person reveals personal or sensitive information to others [21], is a key factor that influences the effectiveness of a mental health chatbot, as it allows chatbots to provide better and more therapeutic experiences to their users [2]. Several factors in chatbots stimulate user self-disclosure, including chatbot accessibility, anonymity, convenience, and perceived non-judgmental nature [11].

However, one of the most frequently mentioned risks of these applications is the privacy and security of user data and their disclosed information [14, 38]. This is particularly sensitive given the risks of stigmatization and discrimination in case the data is disclosed [26, 36]. This is an issue as privacy concerns can lead to mistrust [14, 38], which in turn lead to less disclosure or being a barrier in seeking help [22].

Among various factors that affect user self-disclosure, chatbot modality has received a lot of attention [1, 30, 32]. Voice-based chatbots, in particular, have shown promise in eliciting self-disclosure, with users skipping fewer questions and offering longer and more detailed responses when engaging through speech rather than text [39]. However, this increased verbal disclosure also introduces concerns about identifiability, as users perceive their unique vocal characteristics to be more easily identifiable [29], thus raising potential privacy risks. Given all this, sensitivity of information is a key factor which influences user self-disclosure. Prior work has shown sensitivity to impact user willingness to engage with a topic [10] and it has also been shown that sensitivity is influenced by related privacy concerns [6].

Although prior studies have examined the impact of various factors on self-disclosure such as chatbot modality [39] and the impact of tone [5] and anthropomorphic features [8, 18], little is known about how privacy policies and the sensitivity of questions asked influence mental health self-disclosure as well as whether there is any interaction effect between the two. In this paper, we aim to address this gap by investigating how privacy policies and question sensitivity affects user self-disclosure in a voice-based chatbot. To explore this gap, we propose the following research questions:

Main Research Question

How do privacy policies and the sensitivity of questions impact self-disclosure in a voice-based chatbot?

Research Sub-questions

RQ1 Are users likely to disclose more personal information if they have a better understanding of privacy policy?

RQ2 Does question sensitivity impact the willingness to self-disclose?

RQ3 Is there an interaction effect between user privacy understanding and question sensitivity?

In addressing these research questions, we aim to contribute towards the development of mental health chatbots by investigating the impact of privacy policies and question sensitivity on user self-disclosure. Our findings have implication for future research towards building more transparent, trustworthy, and effective mental health chatbots.

The remainder of the paper is organized as follows. Section 2 reviews relevant work on self-disclosure, the influence of chatbot modalities on self-disclosure, and the importance of privacy policies and research towards making them more concise. Section 3 outlines the user study and the methodology followed in designing it. Section 4 presents the findings of the study. In Section 5, we reflect on the ethical considerations and the reproducibility of our research. Section 6 provides a detailed discussion of the results and highlights the limitations of the study and potential directions for future research. Finally, Section 7 concludes the paper by summarizing our findings, limitations and directions for future work.

2 Background

In this section, we explore prior relevant literature and their contributions. In Section 2.1 we examine research in the domain of mental health chatbots and self-disclosure including factors that influence it. In Section 2.2 we talk about various chatbot modalities and their effects on self-disclosure. In Section 2.3 we examine the importance of privacy policies as well as prior research towards making them more concise and readable. After discussing the literature, we base hypothesis in regards to our own experiment in Section 2.4.

2.1 Self Disclosure with Chatbots

Self-Disclosure is a key factor influencing the capability of mental health chatbots, as it provides the chatbot with more information about the user, which then helps the chatbot respond and provide help to the user. Previous work by Ho et al. [15] has shown that self-disclosing to a chatbot is equally as effective as self-disclosing to a human and can have beneficial emotional, relational, and psychological outcomes. Kahn et al. [17] shows that self-disclosure can reduce stress symptoms and improve positive affect.

As such, a lot of research has been done investigating factors that impact self-disclosure in chatbots. Papneja and Yadav [30] investigate 5 factors that affect self-disclosure - conversational factors, interface modality, user characteristics, mediating mechanisms and contextual factors. They also note that self-disclosure has many dimensions, with a few being the breadth or amount of disclosure, the depth, intimacy, or privacy, the valence and the honesty-accuracy of the information disclosed.

User features such as their age, gender and privacy attitude have also been shown to affect interaction with chatbots and self-disclosure. A user study conducted by Schroeder [32] showed that participants who were more comfortable with technology, younger, and male were more likely to trust the machine. A study by Couper et al. [10] into user willingness to participate in various surveys also showed an effect of sensitivity and general privacy attitude on willingness.

Belen-Saglam et al. [6] investigated the sensitivity of information and its implications for disclosure. Their findings note that privacy concerns are one of the reasons why items are perceived to have a higher sensitivity and that this can consequently affect disclosure. They also identify personal characteristics such as age and gender to influence sensitivity.

In our work, we build on the work of Couper et al. [10] and [6] and investigate the role of sensitivity in self-disclosure particularly in a mental health context as well as its interaction effect with privacy. Additionally, based on the work of Schroeder [32], we account for factors like age, gender, trust in AI and privacy attitude as potential confounding variables.

2.2 Different chatbot modalities

The modality of a chatbot refers to its mode of interaction with users. This can include written, voice-based, and embodied interfaces. A survey conducted by Abd-alrazaq et al. [1] showed that the most common input modality is written language while the most common output modality is a combination of written, spoken and visual languages.

In general, users have been shown to disclose more information when engaging through speech than text [39, 30, 32]. For many users, sharing a dialect with the chatbot is an effective step towards feeling more comfortable during their interactions [14]. Factors such as the tonality of the chatbot [5] and gendered voices [39] have also been shown to affect self-disclosure.

Melzner et al. [29] however do note the risk-benefit trade off that comes with verbal disclosure, namely greater identifiability and greater privacy concerns as users exhibit less self control and provide more affective responses.

In our work, we chose to use a voice-based interface as previous work by Melzner et al. [29] has shown this modality raises increased privacy concerns. Furthermore, we anticipate future mental health applications will increasing adopt voice-based interfaces given that they increase user comfort [14] and increase user self-disclosure Yu et al. [39].

2.3 Importance of Privacy Policies

Privacy policies serve as the primary means through which users are informed of how their data is processed, stored, and shared with other parties. They are especially relevant in a mental health context as users are amongst the most vulnerable of populations and especially at risk of privacy violation via the exploitation of their data [7].

As such, recent work by Lee and Attablayo [20] has shown that people who are more privacy aware tend to disclose more information. If not handled, these privacy concerns can lead to a lack of trust and user's withholding information, which can lead to inaccurate diagnoses and treatment recommendations [31]. This once again reinstates the importance of privacy policies in mental health chatbots.

However, privacy policies are notorious for being extremely long documents and difficult to understand [23, 24, 28]. Wagner [37] shows that the length of the average privacy policy has approximately doubled in the last ten years and quadrupled since 2000. Recent work also shows that the readability of privacy policies has decreased over time [37, 3]. We hypothesize that the failure of many of these policies to be user-friendly, can cause a lack of clarity which creates barriers for users towards trust and self-disclosure.

In order to solve issues caused due to their extensive length and lack of readability, privacy policy summarization techniques have been appealing. Liu et al. [25] proposed a novel abstractive summarization framework that parses source text into a series of semantic graphs before generating the text summary from a summary graph. Tomuro et al. [35] proposed a system for generating summaries of policy statement by categorizing privacy policy sentences into five categories (purpose, third parties, limited collection, limited use and data retention). Sun et al. [34] proposed an large language model (LLM) summarization process that enhances the summary via iterative refinement, through a process of drafting, critiquing and refining the summary.

In our work, we recognize the importance of privacy policies in mental health chatbots shown by Blease et al. [7] and build on the work of Lee and Attablayo [20] by examining how user privacy understanding impacts self-disclosure in a mental health context specifically. To implement this, we use the stepwise prompt chaining technique proposed by Sun et al. [34], to provide users a concise, chatbot specific summary of the privacy policy.

2.4 Hypothesis

Based off of the research detailed in the previous background sections, we draw the following hypothesis -

H1 In regards to **RQ1**, we hypothesize that users shown a privacy policy will have greater privacy understanding and thus self-disclose more information.

H2 In regards to **RQ2**, we hypothesize that as question sensitivity increases, the willingness of the user to self-disclose will decrease.

H3 In regards to **RQ3**, we hypothesize that users with a better understanding of their privacy, will be willing to disclose more information, regardless of sensitivity level.

3 Methodology

In this section, we outline the steps taken to address the research questions outlined earlier (**RQ1**,**RQ2**,**RQ3**). To achieve this, we conducted a mixed design study with two conditions. In the control condition, participants interact with a chatbot interface without knowing its privacy policy. In the experimental condition, participants interact with a chatbot interface that presents an audio recording of the chatbots privacy policy at the start of the interaction. In both conditions, user willingness to self-disclose was measured across 3 question sensitivity levels. In order to gauge user understanding of privacy, we measure user privacy understanding post task for both conditions. This approach ensures we not only measure privacy understanding of users who have heard the privacy policy but also user impressions of the unaddressed privacy policy in the control conditon.

Section 3.1 we explain in detail the experimental design and setup. In Section 3.2 we describe the various variables in our study that we measure. In Section 3.3 we talk about the statistical analysis which will take place after the data collection process.

3.1 Experimental Design

The study consisted of three sequential phases: a pre-task, task, and post-task phase.

Upon choosing to participate and providing their informed consent which can be found in Appendix G, participants were directed to the pre-task phase, hosted on Qualtrics [1]. In this stage, they provided demographic information (age and gender) and responded to questions that assessed their general trust with AI systems adapted from the work of Jian et al. [16]. Their privacy attitude was also evaluated using the Privacy Attitudes Questionnaire (PAQ) introduced by Chignell et al. [9]. A full list of these pre-task questions can be found in Appendix A.

Following the pre-task phase, participants proceeded to the task phase. They were redirected to a web-based mental health chatbot, and randomly assigned one of the two conditions outlined below.

1. **Control**: A chatbot interface without any explanation of its privacy policy. The user interface and mode of interaction can be found illustrated in Figure 1a.
2. **Experimental Condition**: A chatbot interface that included a privacy policy shown at the start of the interaction. The interface and privacy policy are visible in Figure 1b.

The privacy policy in the experimental condition was based off of Woebot[2], a popular mental health chatbot. This approach was taken to ensure that the privacy policy closely resembled those found in real-world applications. We utilize the prompt-chaining technique detailed by Sun et al. [34] to effectively summarize the privacy policy and adapt it to our experiment. Prior to data collection, the privacy policy was piloted to ensure its understandability. The summarized privacy policy which was shown to users can be found in Appendix F.

Participants were then asked to engage with the chatbot across three conversational scenarios covering questions from the following topics. These topics were selected from the work of Ma et al. [27], and were selected for this work as they cover everyday interaction topics which also have a mental health relevance.

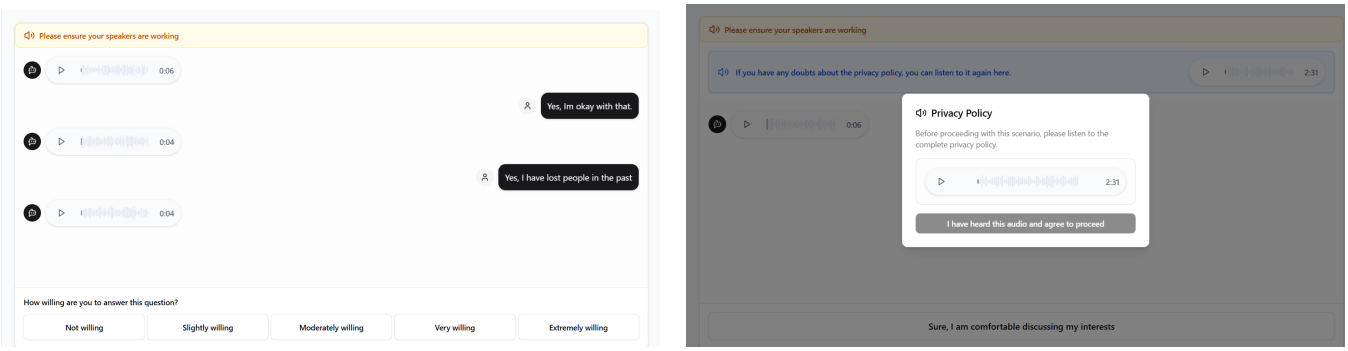
- **Tastes and Interests**
- **Interpersonal Relations and Self-Concept**
- **Work or Studies**

Each scenario consisted of a dialogue comprising of statements and questions. For statements, participants had a single response option which could be selected to continue the dialogue. For each question, participants were asked to rate:

- **Willingness to respond**: Rated on a 5-point Likert scale ranging from 1 (*Not Willing*) to 5 (*Extremely Willing*).

[1] https://www.qualtrics.com/

[2] https://woebothealth.com/



(a) Control Condition

(b) Experimental Condition

Figure 1: Comparison of Control and Experimental Conditions

- **Perceived Sensitivity**: Categorized as *Low*, *Medium*, or *High*.

Each scenario consisted of three questions with varying levels of self-disclosure intimacy. These questions were drawn from the *SelfDisclosureItems* dataset developed by Ma et al. [27]. The order of scenarios was randomized to ensure that there was no impact of order effects.

Finally, after completing the task phase, the participants were redirected to a post-task questionnaire hosted on Qualtrics. This questionnaire collected data on participants' experiences during their interaction with the chatbot, including their understanding of the privacy policy which can be found in Appendix C.1. The questions follow the same approach as Korunovska et al. [19], who evaluated privacy comprehension using a two-option format in which participants selected the option that correctly reflected a data right or threat presented in the privacy policy.

The study was carried out with a total of 26 participants, split evenly between the two conditions. Participant were recruited exclusively from personal networks using snowball sampling. All participant data was anonymized and no remuneration was provided for their participation, so as to ensure there was nothing inducing or biasing participation. A more detailed overview of the breakdown of participants by task can be found in Appendix D.1.

3.2 Variables

There are two independent variables in this study. The first is the between-subjects independent variable which was the presence or absence of a privacy policy (categorical). The second is the within-subjects independent variable which is the question sensitivity (categorical).

The one dependent variable was the self-disclosure willingness, a continuous variable derived from the 5 point Likert scale after each question in the scenarios. The perceived sensitivity of each question was also measured, categorized as *Low*, *Medium*, or *High*. The perceived sensitivity was measured as to explore how perception of sensitivity can vary and consequently affect self-disclosure willingness.

In addition, participant characteristics such as age, gender, trust in AI and privacy attitude were measured before the task as confounds to adjust for their impact on the willingness to disclose. After task completion, the user understanding of the privacy policy was measured to validate its operationalization.

3.3 Statistical Analysis

In our study, we conduct a mixed design study with two independent variables (privacy policy and question sensitivity) and one dependent variable (willingness to self-disclose). We also have 4 confounding variables - age, gender, trust in AI and privacy attitude. For all analysis done in this paper, we use GPower [13] to calculate the required sample size using a priori power analysis and JASP [3] to carry out all the statistical analysis tests detailed later in this section.

We first carry out an independent samples t-test to check for statistical significance between the control and experimental conditions based on participant privacy policy understanding seen in Appendix C.1, which participants attempted post-task.

Given that both the within-subjects independent variable (question sensitivity) and the between-subjects independent variable (privacy policy) in our study were categorical and that our dependent variable (self-disclosure willingness) was continuous, we find the ideal statistical test to analyze our results to be a mixed ANOVA.

The mixed ANOVA displays the impacts of the privacy policy, different question sensitivity levels as well as the interaction effect between privacy policy and question sensitivity. Before carrying out the mixed ANOVA, we test to ensure that its three assumptions hold and that they are not violated. The assumptions and how we test for them are outlined below.

1. **Normality** - We test for normality of data using the Shapiro-Wilk test.

2. **Homogeneity of variance** - We test for homogeneity of variance using Levene's test.

[3] https://jasp-stats.org/

3. **Sphericity** - We test for sphericity using Mauchly's test of sphericity.

In the case of a statistically significant p value being achieved for the between-subjects factor or the within-subjects factor, we run post hoc tests to investigate this significance. Given our small sample size, we also report confidence intervals and effect sizes. Simple Main Effects tests are run to measure the interaction effect. Finally, we extend the mixed ANOVA with covariates to control for our potential confounding variables.

For our mixed ANOVA, we assume a medium effect size of 0.25. Since we are testing for multiple hypothesis, we apply a Bonferroni correction and get a significance level of α = 0.05/3 = 0.01667. Given this, a power of 1-β equal to 0.8, 2 groups (Control and Experiment) with 3 measurements each (Low, Medium and High Sensitivity) and a nonsphericity correction of ϵ = 1, an estimated sample size of 102 participants per group, totaling 204 participants,is required.

Given the short time frame of the project, this however was not possible, and thus the data collection proceeded for a sample size of 26 participants (13 participants per condition). Despite the required sample size not being met, the analysis is conducted as it would have been had the target sample size been achieved, to serve illustrative and demonstrative purposes.

4 Results

This section presents the results of our user study in regards to the main research question and its sub-questions as per the analysis plan. We start by testing whether all the assumptions of mixed ANOVA hold. The assumption test results, which can be found in Appendix D.3, show that all the assumptions hold. Shapiro-Wilk's test shows $p > 0.05$ for all combinations of independent variables proving normality. Levene's test proves homogeneity of variances with $p > 0.05$ for all question sensitivities and Mauchly's test indicated sphericity ($W(2) = 0.906, p = 0.319$).

Having ensured that all assumptions hold, we proceed to carry out the mixed ANOVA. In the following subsections, we will present an overview of the results alongside supporting figures. In Section 4.1, we display the impact of privacy policy on self-disclosure and the user understanding of the privacy policy. In Section 4.2 we present the effects of question sensitivity on self-disclosure. In Section 4.3 we display the interaction effect and extend the mixed ANOVA with covariates.

The final results shown below are based on the data collected from 26 participants (13 per condition). Of the 26 participants, all 26 passed the attention checks present in the pre-task and post-task. The participant age and gender demographics can be seen in Table 1 and Table 2 respectively.

4.1 Privacy Impact on Self Disclosure

To effectively test our operationalization, we evaluate the participants understanding of the privacy policy in both the control and experimental condition using the questions in Appendix C.1. The resulting scores can be seen in Figure 2.

Age Group	Number of Participants
16-20	12
21-25	14
Total	26

Table 1: Participant Age Distribution

Gender	Number of Participants
Male	15
Female	11
Total	26

Table 2: Participant Gender Distribution

As expected, participants in the experimental group show a good understanding of the privacy policy ($\mu = 4.15, \sigma = 0.80$). What is surprising is that, participants in the control condition, who were not shown the privacy policy, also display a good understanding of the privacy policy ($\mu = 3.77, \sigma = 1.09$) . We explore potential reasons for this in Section 6.

An independent samples t-test was carried out to compare the results of the two conditions. The analysis showed that while in general, participants from the experimental condition performed slightly better, there wasn't a statistically significant difference in privacy understanding between the control and experimental condition ($t(24) = -1.024, p = 0.316$).

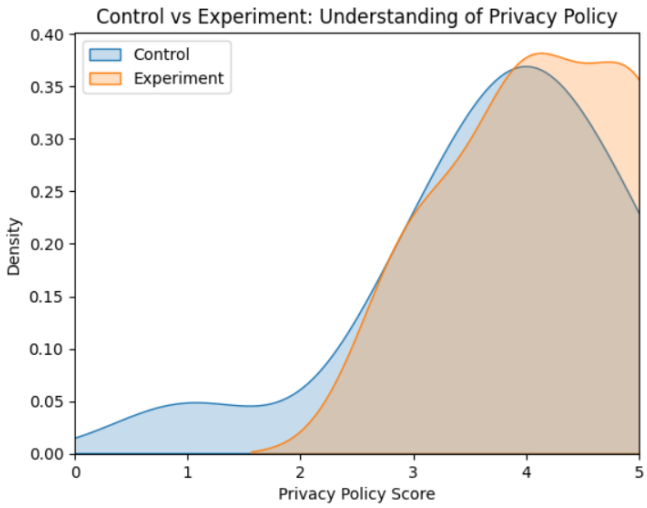


Figure 2: Privacy Policy Understanding

The privacy policy understanding is reflected in the average willingness to self-disclose which is illustrated in Figure 3. The results show similar willingness in both conditions with the experimental conditon ($\mu = 3.62, \sigma = 0.62$) having a higher average willingness and the control conditon ($\mu = 3.48, \sigma = 0.74$) having a higher standard deviation. The results from the between-subjects effects of the mixed ANOVA, seen in Table 12 show that there is no statistical significance ($F(1, 24) = 0.293, p = 0.594$) in the willingness to self-disclose between the control and experimental conditon.

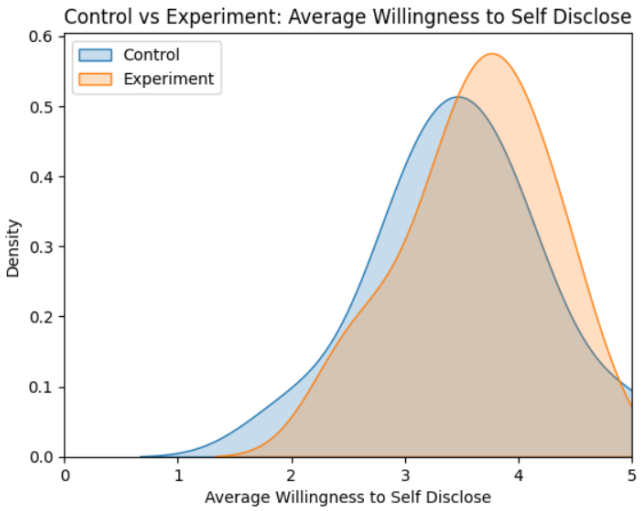


Figure 3: Willingness to Self-Disclose

4.2 Sensitivity Impact on Self Disclosure

We record the impact of question sensitivity on willingness to self-disclose as our within-subjects independent variable. The results can be seen in the box plots shown in Figure 4.

An overall clear trend can be seen between questions of different severities. Participants show similar willingness to self-disclose responses to low sensitivity ($\mu = 3.68, \sigma = 0.77$) and medium sensitivity questions ($\mu = 3.83, \sigma = 0.76$). However, this willingness reduces for high sensitivity questions ($\mu = 3.14, \sigma = 0.75$).

Our mixed ANOVA results for within-subjects effects, seen in Table 11 show a statistical significance for sensitivity ($F(2, 48) = 18.614, p < 0.01, \omega^2 = 0.128$) showing a medium to large effect . Thus we carry out post hoc tests, which can be found in Appendix D.5 to examine this further using a confidence interval (CI) of 95% and Cohen's d for reporting effect sizes. Our post hoc tests display a statistical significance in willingness between low and high ($M = 0.538, SE = 0.113, P < 0.01$) with a medium to large effect ($d = 0.700, CI[0.241, 1.158]$) . There is also a statistical significance between medium and high sensitivity questions ($M = 0.692, SE = 0.107, P < 0.01$) with a large effect ($d = 0.900, CI[0.411, 1.388]$) . It is also noted that there is no significant difference between low and medium sensitivity questions ($M = -0.154, SE = 0.136, P = 0.807$).

We also display a comparison of self-disclosure willingness by question sensitivity and disclosure willingness by perceived sensitivity to explore whether question sensitivity is perceived differently between people. The results can be seen in the box plots in Appendix D.8.

4.3 Interaction effect and Covariates

After examining the individual effects of the between-subjects and within-subjects factors, we examine their interaction effect. A descriptive overview of these results can be seen in the box plots in Appendix D.7.

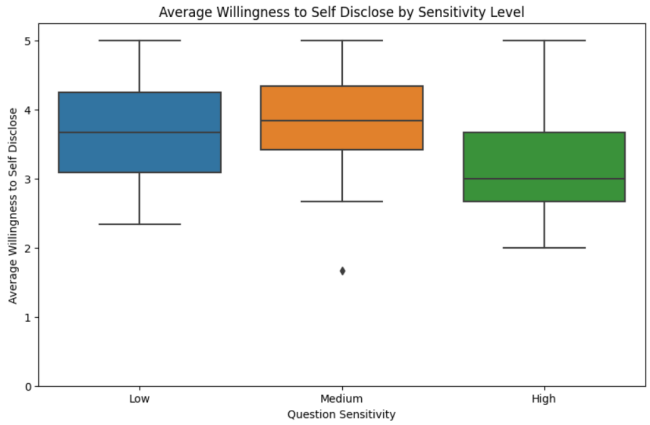


Figure 4: Question Sensitivity based Willingness

The box plots show similar statistics in willingness across conditions for a particular sensitivity. This is supported by the simple main effects test shows which there is no interaction effect for low ($F(1, 48) = 0.063, p = 0.804$), medium ($F(1, 48) = 0.357, p = 0.556$) or high sensitivity ($F(1, 48) = 0.363, p = 0.552$) with privacy policy.

Analysis was also done factoring in the covariates age, gender, privacy attitude and trust in AI. The results can be found in Appendix D.9. Analysis with the addition of the covariates suggests that only privacy attitude had a significant effect ($F(1, 32) = 8.190, p = 0.011$) between the control and experimental conditions and resulted in a more significant p value ($p = 0.354$). Question sensitivity was initially significant ($F(2, 48) = 18.614, p < 0.01, \omega^2 = 0.128$) but on adding the covariates becomes insignificant ($F(2, 32) = 18.614, p = 0.648, \omega^2 = 0.00$). The significance of the interaction effect is shown to increase for low ($F(1, 32) = 0.168, p = 0.687$), medium ($F(1, 32) = 0.751, p = 0.399$) or high sensitivity ($F(1, 32) = 1.561, p = 0.229$) with the addition of covariates.

5 Responsible Research

This section outlines the work done to ensure that the research is done responsibly, in accordance with ethics, and in a reproducible manner in the spirit of transparent research and with integrity.

Before the study was carried out, an application was submitted and approved by the TU Delft Human Research Ethics Committee (HREC) under ethics ID 5399. The HREC application consisted of a risk assessment and mitigation plan surrounding potential risks involving participants, their recruitment, as well as data protection and privacy. Mitigation strategies including informed consent and not collecting personally identifiable research data are described below.

Along with the HREC, a data management plan (DMP) was also approved. The DMP outlined data collection, storage, use and documentation and how it would be carried out in an ethical and legal manner. All data was stored in a GDPR compliant server and will later be stored anonymously in SURF-

Drive.

No Personally Identifiable Research Data (PIRD) was collected. Only minimal demographic data, age (in five-year bins), and gender were collected, so as to ensure participants remained anonymous and were not identifiable based on the data they provided.

We also identified disclosure to be of a sensitive nature and thus to minimize any emotional discomfort, participants were only asked about their willingness to disclose sensitive information and not to disclose the information itself. The recruitment for the study and study were also designed to ensure participation was voluntary and that participants could withdraw at any point of time.

When starting their participation in the study, participants were briefed on the purpose, handling of data, and their rights in the study through an informed consent form which can be found in the appendix G. The study continued only once consent had been given. Participants also had the option to withdraw at any time without consequence at which point all their data collected up to that point was voided.

Finally, to support transparency and reproducibility, all code related to this project has been made publicly available and questionnaires can all be found in the Appendix. All data used was in keeping with the principles of FAIR (Findability, Accesibility, Interoperability and Reusability) to maximize its utility for future research. Furthermore, anonymized datasets will be made publicly available upon publication, and limitations are openly discussed in Section 6.2.

In the interest of full transparency and integrity, we report that large language models (LLMs) were used during the writing process to generate complex formats of tables and figures used as well as to improve their alignment and positioning. LLMs were not used for any idea generation or result analysis done in this report. All outputs from LLMs were carefully reviewed and verified before including them in this report. Example prompts used for the purposes outlines above can be found in Appendix E.

6 Discussion

In this section, we discuss the implications of the results presented in Section 4 as well as possible limitations of our study and potential directions for future research.

6.1 Interpretation of Results

Our study investigated the impact of privacy policies and sensitivity of questions on willingness of users to self-disclose information as well as their interaction effect in a mental health chatbot.

As expected, participants in the experimental condition reported a high understanding of the privacy policy. Surprisingly, participants in the control conditon represented a good understanding of the privacy policy of the chatbot, despite not being explicitly shown it. This unexpected outcome suggests that despite not having explicit exposure, participants of the control condition had pre-existing notions and general awareness of privacy which aligned with the privacy policy of the chatbot. This could stem from the fact that most of the participants were below the age of 30 and previous work by Dommeyer and Gross [12] has shown younger people to have a higher privacy awareness as compared to older people. However, it is possible that a more comprehensive test of privacy understanding could yield different results and this should be a point of focus for future research which could for instance evaluate comprehension by having users design a downstream task based on their privacy comprehension.

Similarly, in regards to **RQ1**, participants in the experimental conditon reported a slightly higher willingness to self-disclose. This minimal difference however, aligns with the understanding of privacy policy reported. In our hypothesis **H1**, we expect participants, with a better understanding of the privacy policy to show a greater willingness to disclose. This hypothesis is partially shown to hold, as participants in the experimental conditon reported slightly higher understandings and willingness to self-disclose despite there being no statistical significance.

In regards to **RQ2**, we note that question sensitivity does impact self-disclosure willingness. Our results show a clear trend in self-disclosure willingness with a significantly lower willingness for high sensitivity questions as compared to low and medium sensitivity questions. This is supported by the statistical significance shown in our post hoc tests ($p < 0.01$). This aligns with our hypothesis **H2**, showing that as the sensitivity of a question increases, the self-disclosure willingness decreases. This supports previous work by Couper et al. [10] that users willingness is negatively affected by sensitivity. Further analysis between objective question sensitivity and perceived question sensitivity also indicated more distinct trends in willingness to disclose for perceived sensitivity. This finding suggests a role of perception towards disclosure and could be a promising domain for future research.

In regards to **RQ3**, we note that there is no interaction effect between privacy policy and question sensitivity with willingness being similar across conditions for each sensitivity of questions. However this is expected, given that the privacy policy understanding is very similar in both conditions and thus willingness across conditions for different sensitivities also remains similar. This partially aligns with our hypothesis **H3** which states that higher privacy understanding will increase self-disclosure willingness across all question sensitivities.

6.2 Limitations and Future Work

There are multiple limitations which must be taken into account in regards to the study. This however, also opens up avenues for future research.

Firstly, participant recruitment from a personal network resulted in demographics that may not accurately represent individuals with mental health concerns, who are the intended users of such applications. Consequently, the behavior of participants when interacting with the system could vary significantly from that of the target user group. Future research

should aim for a more diverse and representative participant pool, potentially through targeted recruitment that also reaches individuals actively using similar applications.

Secondly, the study followed a very guided form of interaction where participants were unable to communicate in free form text and did not self-disclose any information but merely recorded their willingness to self-disclose. This structured approach may lead to discrepancies with a real-life interaction where users actually disclose information. Future work should investigate self-disclosure within more free-form communication within a mental health chatbot. Additionally, exploring various dimensions of self-disclosure such as the depth and intimacy, as noted in previous research, could provide a more comprehensive understanding of user behavior.

Due to time constraints, it was not possible to carry out the study with the required participant size in order to establish a statistical significance to the results obtained. Consequently, the findings and results could vary considerably over larger participant pools. Future research should aim to address this by carrying out studies with larger and statistically robust sample sizes to allow for more conclusive findings.

Lastly, surveys limit the amount of understanding that can be gained regarding user behavior, Future research could explore semi-structured interviews to delve deeper into user perceptions, allowing for in-depth discussions that reveal underlying motivations, concerns, and interpretations that surveys might miss.

7 Conclusions and Future Work

In this paper we investigate the impact of privacy policies and question sensitivity on user willingness to self-disclose when interacting with a voice based mental health chatbots. We explore whether a privacy policy leads to higher self-disclosure, the influence of question sensitivity on self-disclosure and if there is any interaction effect between the two factors. To this end, we conducted a mixed design study with 26 participants.

Our findings demonstrate that simply showing a privacy policy did not result in a significant difference in privacy understanding and thus willingness to disclose. We identify however that willingness to self-disclose decreased as question sensitivity increased, particularly between low, medium sensitivity questions and high sensitivity questions. Our findings suggest that simply providing a privacy policy might not be sufficient to address user privacy concerns or improve willingness to disclose.

For future work, it is imperative to conduct studies with larger and more representative participant pools to achieve statistical significance. Furthermore, a deeper analysis of how individual differences in privacy attitudes and trust in AI interact with privacy explanations to influence self-disclosure is warranted. These directions will be crucial for the development of more transparent, trustworthy, and effective mental health chatbots that encourage user self-disclosure while safeguarding user privacy.

References

[1] A. A. Abd-alrazaq, M. Alajlani, A. A. Alalwan, B. M. Bewick, P. Gardner, and M. Househ. An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics*, 132:103978, Dec. 2019. ISSN 1386-5056. doi: 10.1016/j.ijmedinf.2019.103978. URL http://dx.doi.org/10.1016/j.ijmedinf.2019.103978.

[2] A. A. Abd-Alrazaq, M. Alajlani, N. Ali, K. Denecke, B. M. Bewick, and M. Househ. Perceptions and opinions of patients about mental health chatbots: Scoping review. *Journal of Medical Internet Research*, 23(1): e17828, Jan. 2021. ISSN 1438-8871. doi: 10.2196/17828. URL http://dx.doi.org/10.2196/17828.

[3] R. Amos, G. Acar, E. Lucherini, M. Kshirsagar, A. Narayanan, and J. Mayer. Privacy policies over time: Curation and analysis of a million-document dataset. In *Proceedings of the Web Conference 2021*, WWW '21, page 2165–2176, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3450048. URL https://doi-org.tudelft.idm.oclc.org/10.1145/3442381.3450048.

[4] E. Anthes. Mental health: There's an app for that. *Nature*, 532(7597):20–23, Apr. 2016. ISSN 1476-4687. doi: 10.1038/532020a. URL http://dx.doi.org/10.1038/532020a.

[5] S. A. B. S. Baharin, V. Lamarche, N. Weinstein, and S. Paulmann. Interested-sounding voices influence listeners' willingness to disclose. In *Speech Prosody 2024*, ISCA, July 2024. ISCA.

[6] R. Belen-Saglam, J. R. C. Nurse, and D. Hodges. An investigation into the sensitivity of personal information and implications for disclosure: A uk perspective. *Frontiers in Computer Science*, Volume 4 - 2022, 2022. ISSN 2624-9898. doi: 10.3389/fcomp.2022.908245. URL https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2022.908245.

[7] C. Blease, A. Kharko, M. Annoni, J. Gaab, and C. Locher. Machine learning in clinical psychology and psychotherapy education: A mixed methods pilot survey of postgraduate students at a swiss university. *Front. Public Health*, 9:623088, Apr. 2021.

[8] J. Chen, M. Li, and J. Ham. Different dimensions of anthropomorphic design cues: How visual appearance and conversational style influence users' information disclosure tendency towards chatbots. *International Journal of Human Computer Studies*, 190, Oct. 2024. ISSN 1071-5819. doi: 10.1016/j.ijhcs.2024.103320. Publisher Copyright: © 2024 Elsevier Ltd.

[9] M. Chignell, J. Gwizdka, and A. Quan-Haase. The privacy attitudes questionnaire (paq): Initial development and validation. volume 47, 10 2003. doi: 10.1177/154193120304701102.

[10] M. P. Couper, E. Singer, F. G. Conrad, and R. M. Groves. Risk of disclosure, perceptions of risk, and concerns about privacy and confidentiality as factors in survey participation. *J. Off. Stat.*, 24(2):255–275, 2008.

[11] E. A. J. Croes, M. L. Antheunis, C. van der Lee, and J. M. S. de Wit. Digital confessions: The willingness to disclose intimate information to a chatbot and its impact on emotional well-being. *Interacting with Computers*, 36(5):279–292, June 2024. ISSN 1873-7951. doi: 10.1093/iwc/iwae016. URL http://dx.doi.org/10.1093/iwc/iwae016.

[12] C. J. Dommeyer and B. L. Gross. What consumers know and what they do: An investigation of consumer knowledge, awareness, and use of privacy protection strategies. *Journal of Interactive Marketing*, 17(2):34–51, 2003. ISSN 1094-9968. doi: https://doi.org/10.1002/dir.10053. URL https://www.sciencedirect.com/science/article/pii/S1094996803701339.

[13] F. Faul, E. Erdfelder, A. Buchner, and A.-G. Lang. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behav. Res. Methods*, 41(4):1149–1160, Nov. 2009.

[14] M. D. R. Haque and S. Rubya. An overview of chatbot-based mobile mental health apps: Insights from app description and user reviews. *JMIR Mhealth Uhealth*, 11: e44838, May 2023. ISSN 2291-5222. doi: 10.2196/44838. URL https://mhealth.jmir.org/2023/1/e44838.

[15] A. Ho, J. Hancock, and A. S. Miner. Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. *Journal of Communication*, 68(4):712–733, 05 2018. ISSN 0021-9916. doi: 10.1093/joc/jqy026. URL https://doi.org/10.1093/joc/jqy026.

[16] J.-Y. Jian, A. Bisantz, and C. Drury. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4: 53–71, 03 2000. doi: 10.1207/S15327566IJCE0401_04.

[17] J. Kahn, J. Achter, and E. Shambaugh. Client distress disclosure, characteristics at intake, and outcome in brief counseling. *Journal of Counseling Psychology*, 48: 203–211, 04 2001. doi: 10.1037/0022-0167.48.2.203.

[18] E. Kang and Y. Kang. Counseling chatbot design: The effect of anthropomorphic chatbot characteristics on user self-disclosure and companionship. *International Journal of Human-Computer Interaction*, 40(11): 2781–2795, 2024. ISSN 1044-7318. doi: 10.1080/10447318.2022.2163775. Publisher Copyright: © 2023 Taylor Francis Group, LLC.

[19] J. Korunovska, B. Kamleitner, and S. Spiekermann. The challenges and impact of privacy policy comprehension, 05 2020.

[20] K. Lee and P. Attablayo. Examining the impacts of privacy awareness on user's self-disclosure on social media, 2023. URL https://arxiv.org/abs/2303.07927.

[21] Y.-C. Lee, N. Yamashita, Y. Huang, and W. Fu. "i hear you, i feel you": Encouraging deep self-disclosure through a chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376175. URL https://doi.org/10.1145/3313831.3376175.

[22] L. Li, W. Peng, and M. M. Rheu. Factors predicting intentions of adoption and continued use of artificial intelligence chatbots for mental health: Examining the role of utaut model, stigma, privacy concerns, and artificial intelligence hesitancy. *Telemedicine and e-Health*, 30(3):722–730, 2024. doi: 10.1089/tmj.2023.0313. URL https://doi.org/10.1089/tmj.2023.0313. PMID: 37756224.

[23] T. Libert. An automated approach to auditing disclosure of third-party data collection in website privacy policies. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 207–216, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee. ISBN 9781450356398. doi: 10.1145/3178876.3186087. URL https://doi-org.tudelft.idm.oclc.org/10.1145/3178876.3186087.

[24] T. Linden, R. Khandelwal, H. Harkous, and K. Fawaz. The privacy policy landscape after the gdpr, 2019. URL https://arxiv.org/abs/1809.08396.

[25] F. Liu, J. Flanigan, S. Thomson, N. Sadeh, and N. A. Smith. Toward abstractive summarization using semantic representations, 2018. URL https://arxiv.org/abs/1805.10399.

[26] C. A. Lovejoy. Technology and mental health: The role of artificial intelligence. *European Psychiatry*, 55:1–3, 2019. doi: 10.1016/j.eurpsy.2018.08.004.

[27] X. Ma, J. Hancock, and M. Naaman. Anonymity, intimacy and self-disclosure in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 3857–3869, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450333627. doi: 10.1145/2858036.2858414. URL https://doi.org/10.1145/2858036.2858414.

[28] A. M. McDonald and L. F. Cranor. The cost of reading privacy policies. *I/S: A Journal of Law and Policy for the Information Society*, 4:543–568, 2008.

[29] J. Melzner, A. Bonezzi, and T. Meyvis. Information disclosure in the era of voice technology. *Journal of Marketing*, 87(4):491–509, Mar. 2023. ISSN 1547-7185. doi: 10.1177/00222429221138286. URL http://dx.doi.org/10.1177/00222429221138286.

[30] H. Papneja and N. Yadav. Self-disclosure to conversational ai: a literature review, emergent framework, and directions for future research. *Personal and Ubiquitous Computing*, 29(2):119–151, Aug. 2024. ISSN 1617-4917. doi: 10.1007/s00779-024-01823-7. URL http://dx.doi.org/10.1007/s00779-024-01823-7.

[31] M. Rahsepar Meadi, T. Sillekens, S. Metselaar, A. van Balkom, J. Bernstein, and N. Batelaan. Exploring the ethical challenges of conversational AI in mental health care: Scoping review. *JMIR Ment. Health*, 12:e60432, Feb. 2025.

[32] J. Schroeder. Trusting in machines: How mode of interaction affects willingness to share personal information with machines. In *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018. URL https://scholarspace.manoa.hawaii.edu/items/903c4cf7-3335-4bc8-8436-8e34c379c8c5.

[33] Z. Steel, C. Marnane, C. Iranpour, T. Chey, J. W. Jackson, V. Patel, and D. Silove. The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013. *International Journal of Epidemiology*, 43(2):476–493, Mar. 2014. ISSN 0300-5771. doi: 10.1093/ije/dyu038. URL http://dx.doi.org/10.1093/ije/dyu038.

[34] S. Sun, R. Yuan, Z. Cao, W. Li, and P. Liu. Prompt chaining or stepwise prompt? refinement in text summarization, 2024. URL https://arxiv.org/abs/2406.00507.

[35] N. Tomuro, S. Lytinen, and K. Hornsburg. Automatic summarization of privacy policies using ensemble learning. In *Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy*, CODASPY '16, page 133–135, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450339353. doi: 10.1145/2857705.2857741. URL https://doi-org.tudelft.idm.oclc.org/10.1145/2857705.2857741.

[36] G. N. Vilaza and D. McCashin. Is the automation of digital mental health ethical? applying an ethical framework to chatbots for cognitive behaviour therapy. *Front. Digit. Health*, 3:689736, Aug. 2021.

[37] I. Wagner. Privacy policies across the ages: Content and readability of privacy policies 1996–2021, 2022. URL https://arxiv.org/abs/2201.08739.

[38] B. Wies, C. Landers, and M. Ienca. Digital mental health for young people: A scoping review of ethical promises and challenges. *Front. Digit. Health*, 3:697072, Sept. 2021.

[39] Q. Yu, T. Nguyen, S. Prakkamakul, and N. Salehi. "i almost fell in love with a machine": Speaking with computers affects self-disclosure. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–6. ACM, May 2019. doi: 10.1145/3290607.3312918. URL http://dx.doi.org/10.1145/3290607.3312918.

A Pre-task questions

A.1 General Trust in AI

Statement	Strongly disagree	Disagree	Neither agree or disagree	Agree	Strongly Agree
I trust AI systems to operate reliably.	○	○	○	○	○
I feel comfortable relying on AI systems to make decisions.	○	○	○	○	○
I believe AI systems can be depended on.	○	○	○	○	○
I am cautious when using AI systems.	○	○	○	○	○
AI systems are trustworthy in most situations.	○	○	○	○	○

Table 3: General Trust in AI Questionaire

A.2 Privacy Attitude

Statement	Strongly disagree	Disagree	Neither agree or disagree	Agree	Strongly Agree
I am comfortable with giving a DNA sample.	○	○	○	○	○
I am comfortable giving out my personal identification number.	○	○	○	○	○
I am comfortable in allowing others to check my credit.	○	○	○	○	○
I am comfortable wearing a name tag.	○	○	○	○	○
Employers should be able to monitor employee email.	○	○	○	○	○
It is ok to use messaging services even if the messages could in principle be tracked.	○	○	○	○	○
I allow strangers to enter my house while I'm not there.	○	○	○	○	○
I am comfortable with having my retina scanned.	○	○	○	○	○
I do not mind using my real name in online discussions.	○	○	○	○	○
My medical information should never be communicated to people or organizations without my permission.	○	○	○	○	○

Table 4: Privacy Attitude Questionaire

B Task

All source code relating to the task phase of the user study can be found in this mhealth-chatbot github [4]. This includes code describing the implementation of the chatbot interface as well as the different scenarios participants navigated through and the text to speech conversion process using Google Text To Speech (GTTS) [5]. All scenarios used in the study can be found in the SelfDisclosureItems dataset [6].

C Post-task questions

C.1 Privacy Policy Understanding

Statement	True	False
This study is not compliant with GDPR regulations.	○	○
The only data collected is willingness to self-diclose and perceived sensitivity.	○	○
Data will later be stored anonymously on Surf Drive.	○	○
Users may revoke consent at any time and have all data collected voided.	○	○
There is no contact person in case of any personal data concerns.	○	○

Table 5: Privacy Policy Understanding

D Analysis

D.1 Participant Demographics

Task Condition	Gender	Age		Total
		16–20	21–25	
1	Female	4	1	5
	Male	4	4	8
	Total	8	5	13
2	Female	2	4	6
	Male	2	5	7
	Total	4	9	13
Total	Female	6	5	11
	Male	6	9	15
	Total	12	14	26

Table 6: Participant Demographics

[4] https://github.com/Sagar-CK/mhealth-chatbot
[5] https://pypi.org/project/gTTS/
[6] https://github.com/sTechLab/SelfDisclosureItems

D.2 Independent Samples T-Test

	t	df	p
Privacy Policy Understanding	-1.024	24	0.316

Table 7: Independent Samples T-Test

D.3 Mixed ANOVA Assumption Tests

	avg_low_sensitivity		avg_medium_sensitivity		avg_high_sensitivity	
	1	2	1	2	1	2
Valid	13	13	13	13	13	13
Missing	0	0	0	0	0	0
Mean	3.641	3.718	3.744	3.923	3.051	3.231
Std. Deviation	0.799	0.768	0.884	0.626	0.859	0.644
Shapiro-Wilk	0.976	0.970	0.940	0.974	0.921	0.931
P-value of Shapiro-Wilk	0.958	0.899	0.457	0.938	0.261	0.356
Minimum	2.333	2.333	1.667	2.667	2.000	2.000
Maximum	5.000	5.000	5.000	5.000	5.000	4.000

Table 8: Descriptive Statistics and Shapiro-Wilk Test

	F	df1	df2	p
avg_low_sensitivity	0.035	1	24	0.854
avg_medium_sensitivity	0.675	1	24	0.419
avg_high_sensitivity	0.899	1	24	0.352

Table 9: Levene's Test

	Mauchly's W	Approx. χ^2	df	p-value	Greenhouse-Geisser ϵ	Huynh-Feldt ϵ	Lower Bound ϵ
Sensitivity	0.906	2.282	2	0.319	0.914	0.985	0.500

Table 10: Mauchly's Sphericity Test

D.4 Mixed ANOVA Results

Cases	Sphericity Correction	Sum of Squares	df	Mean Square	F	p	ω^2
Sensitivity	None	6.872	2.000	3.436	18.614	$< .001$	0.128
	Greenhouse-Geisser	6.872	1.827	3.760	18.614	$< .001$	0.128
	Huynh-Feldt	6.872	1.970	3.488	18.614	$< .001$	0.128
Sensitivity × condition_task	None	0.046	2.000	0.023	0.123	0.884	0.000
	Greenhouse-Geisser	0.046	1.827	0.025	0.123	0.867	0.000
	Huynh-Feldt	0.046	1.970	0.023	0.123	0.881	0.000
Residuals	None	8.860	48.000	0.185			
	Greenhouse-Geisser	8.860	43.857	0.202			
	Huynh-Feldt	8.860	47.285	0.187			

Note. Type III Sum of Squares

Table 11: Within Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	ω^2
condition_task	0.412	1	0.412	0.293	0.594	0.000
Residuals	33.772	24	1.407			

Note. Type III Sum of Squares

Table 12: Between Subjects Effects

D.5 POST Hoc Tests (Mixed ANOVA)

		95% CI for Mean Difference						95% CI for Cohen's d			
		Mean Difference	Lower	Upper	SE	df	t	Cohen's d	Lower	Upper	P_{Bonf}
Low	Medium	-0.154	-0.504	0.196	0.136	24	-1.131	-0.200	-0.661	0.261	0.807
	High	0.538	0.248	0.829	0.113	24	4.771	0.700	0.241	1.158	$< .001^{***}$
Medium	High	0.692	0.418	0.967	0.107	24	6.493	0.900	0.411	1.388	$< .001^{***}$

*** $p < .001$

Note. P-value and confidence intervals adjusted for comparing a family of 3 estimates (confidence intervals corrected using the bonferroni method).
Note. Results are averaged over the levels of: condition, task

Table 13: post hoc sensitivity test

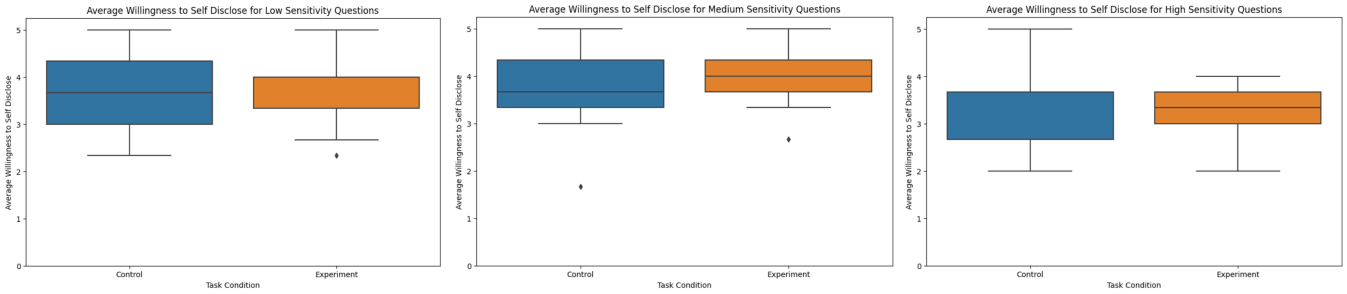
D.6 Simple Main Effects

Level of Sensitivity	Sum of Squares	df	Mean Square	F	p
Low	0.038	1	0.038	0.063	0.804
Medium	0.209	1	0.209	0.357	0.556
High	0.209	1	0.209	0.363	0.552

Note. Type III Sum of Squares

Table 14: Simple Main Effects

D.7 Interaction Effect - Sensitivity and Privacy



(a) Low Sensitivity (b) Medium Sensitivity (c) High Sensitivity

Figure 5: Interaction effect between question sensitivity and privacy policy

D.8 Descriptive Analysis of Perceived and Question Sensitivity

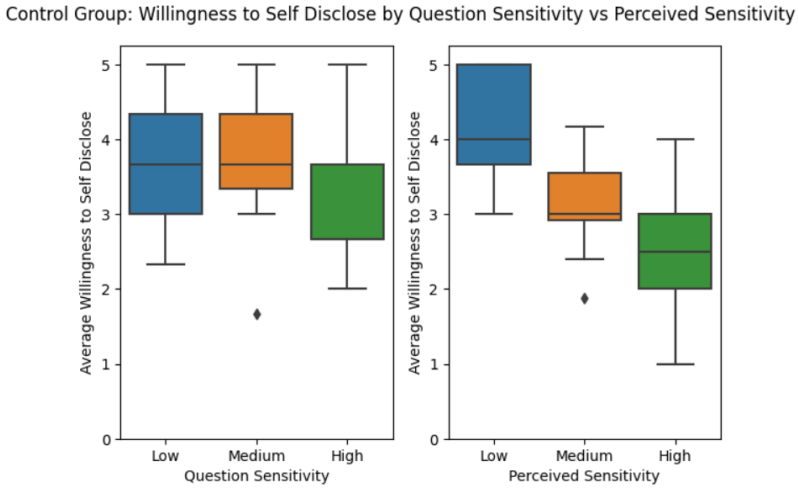


Figure 6: Control Group: Willingness to Self Disclose by Question Sensitivity vs Perceived Sensitivity

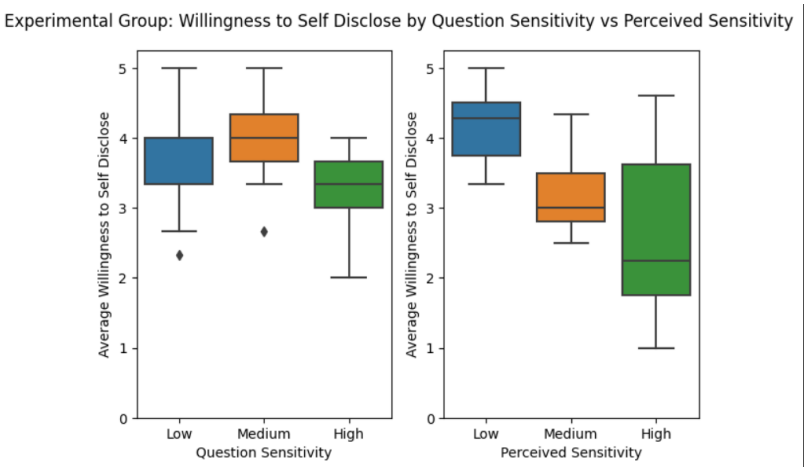


Figure 7: Experimental Group: Willingness to Self Disclose by Question Sensitivity vs Perceived Sensitivity

D.9 Mixed ANOVA with covariates

Cases	Sphericity Correction	Sum of Squares	df	Mean Square	F	p	ω^2
Sensitivity	None	0.163	2.000	0.081	0.439	0.648	0.000
	Greenhouse-Geisser	0.163	1.888	0.086	0.439	0.616	0.000
	Huynh-Feldt	0.163	1.865	0.087	0.439	0.635	0.000
Sensitivity × condition_task	None	0.149	2.000	0.074	0.402	0.673	0.000
	Greenhouse-Geisser	0.149	1.688	0.088	0.402	0.639	0.000
	Huynh-Feldt	0.149	1.865	0.080	0.402	0.659	0.000
Sensitivity × privacy_attitude	None	0.283	2.000	0.142	0.764	0.474	0.000
	Greenhouse-Geisser	0.283	1.688	0.168	0.764	0.455	0.000
	Huynh-Feldt	0.283	1.865	0.152	0.764	0.466	0.000
Sensitivity × trust_in_AI	None	0.061	2.000	0.031	0.165	0.849	0.000
	Greenhouse-Geisser	0.061	1.688	0.036	0.165	0.813	0.000
	Huynh-Feldt	0.061	1.865	0.033	0.165	0.835	0.000
Sensitivity × Gender	None	0.750	2.000	0.375	2.024	0.149	0.015
	Greenhouse-Geisser	0.750	1.688	0.445	2.024	0.157	0.015
	Huynh-Feldt	0.750	1.865	0.402	2.024	0.152	0.015
Sensitivity × Age	None	0.808	2.000	0.404	2.180	0.130	0.017
	Greenhouse-Geisser	0.808	1.688	0.479	2.180	0.139	0.017
	Huynh-Feldt	0.808	1.865	0.433	2.180	0.134	0.017
Sensitivity × condition_task × Gender	None	0.033	2.000	0.016	0.088	0.916	0.000
	Greenhouse-Geisser	0.033	1.688	0.019	0.088	0.886	0.000
	Huynh-Feldt	0.033	1.865	0.018	0.088	0.904	0.000
Sensitivity × condition_task × Age	None	0.375	2.000	0.187	1.011	0.375	1.596×10^{-4}
	Greenhouse-Geisser	0.375	1.688	0.222	1.011	0.365	1.596×10^{-4}
	Huynh-Feldt	0.375	1.865	0.201	1.011	0.371	1.596×10^{-4}
Sensitivity × Gender × Age	None	0.095	2.000	0.047	0.255	0.776	0.000
	Greenhouse-Geisser	0.095	1.688	0.056	0.255	0.739	0.000
	Huynh-Feldt	0.095	1.865	0.051	0.255	0.761	0.000
Sensitivity × condition_task × Gender × Age	None	0.337	2.000	0.168	0.908	0.413	0.000
	Greenhouse-Geisser	0.337	1.688	0.200	0.908	0.400	0.000
	Huynh-Feldt	0.337	1.865	0.181	0.908	0.408	0.000
Residuals	None	5.931	32.000	0.185			
	Greenhouse-Geisser	5.931	27.003	0.220			
	Huynh-Feldt	5.931	29.839	0.199			

Note. Type III Sum of Squares

Table 15: Within Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	ω^2
condition_task	0.985	1	0.985	0.910	0.354	0.000
privacy_attitude	8.863	1	8.863	8.190	0.011	0.175
trust_in_AI	0.746	1	0.746	0.689	0.419	0.000
Gender	0.557	1	0.557	0.515	0.483	0.000
Age	0.246	1	0.246	0.227	0.640	0.000
condition_task × Gender	4.393	1	4.393	4.059	0.061	0.083
condition_task × Age	3.490	1	3.490	3.225	0.091	0.061
Gender × Age	0.011	1	0.011	0.010	0.920	0.000
condition_task × Gender × Age	0.897	1	0.897	0.829	0.376	0.000
Residuals	17.314	16	1.082			

Note. Type III Sum of Squares

Table 16: Between Subjects Effects

Level of Sensitivity	Sum of Squares	df	Mean Square	F	p
Low	0.097	1	0.097	0.168	0.687
Medium	0.304	1	0.304	0.751	0.399
High	0.733	1	0.733	1.561	0.229

Note. Type III Sum of Squares

Table 17: Simple Main Effects

E LLM prompts

LLMs were utilized during the writing of this report mainly for the purpose of figure and table formatting in LaTeXsuch as using multi-columns, mini-pages and making images and tables better aligned and positioned. Example prompts of this usage are mentioned below.

- "Can you provide me LaTeXcode that matches the format of this table"
- "Can you make this code such that the images are side by side"

F Privacy Policy

We are committed to safeguarding personal data in compliance with international privacy standards such as the General Data Protection Regulation (GDPR). We ensure transparency in our data practices, guarantee that personal data is never sold to advertisers, and commit to storing data anonymously whenever possible.

We collect personal data such as users' willingness to self-disclose and their perceived sensitivity of questions. This data is gathered with specific intent—to support academic research in developing mobile health (mHealth) chatbots and to personalize and improve the user experience with the Services.

Personal data may be shared exclusively for academic and research purposes. All shared data is de-identified or anonymized to ensure that individual user identities remain protected.

We use GDPR-compliant data servers to securely store personal data. Collected data is later transferred and stored anonymously on Surf Drive or TU Delft's secure storage infrastructure.

Personal data is retained only for as long as needed to fulfill the purposes for which it was collected, including supporting research and delivering services. Users may revoke their consent at any time, at which point all collected data will be voided and deleted in accordance with applicable regulations.

Users have the right to access, correct, and delete their data. They may also revoke consent, restrict processing, and request a copy of their data, as outlined by data protection laws.

Users can withdraw consent at any time, which will result in the deletion and invalidation of all previously collected data.

Users are encouraged to regularly review this privacy policy to stay informed about how their data is handled. For questions or concerns regarding personal data, users can contact e.c.s.degroot@tudelft.nl.

Please note that this policy is in regards to the chatbot interface specifically and interactions with it.

G Informed Consent

Thank you for your interest in participating in our study. This study is led by researchers from the Delft University of Technology and is part of a bachelor thesis conducted by Lina Sadoukri, Yushan Shan, Sagar Chethan Kumar and Manu Gautam.

The purpose of this research study is to investigate factors that relate to the willingness to disclose information to a mental health application. The study will take approximately 5 8 minutes to complete. The data will be used for scientific and educational purposes and may result in a scientific publication.

As part of this study, you will interact with a mental health chatbot. You will receive questions related to you, your mental health, and your well-being. We will not ask you to answer these questions, but rather to indicate how willing you would be to answer them. Additionally, we will ask you about your gender, age, and your agreement with certain statements (e.g., attitudes towards technology) through pre-task and post-task surveys. There are no right or wrong answers.

As with any online activity, there is a potential risk of data breach. We will minimize this risk by not collecting your

name, contact details, or IP address. All data collected will be fully anonymous and cannot be traced back to you. Anonymous data may be publicly shared for scientific purposes.

Your participation is completely voluntary. If you do not complete your submission, your data will not be stored and your participation will be considered withdrawn.

If you have any questions or wish to omit any responses, please contact the responsible researcher:

Esra de Groot, e.c.s.degroot@tudelft.nl

- ○ I consent, begin the study
- ○ I do not consent, I do not wish to participate