Taylor & Francis
Taylor & Francis Group

Journal of Health Communication
International Perspectives

ISSN: 1081-0730 (Print) 1087-0415 (Online) Journal homepage: www.tandfonline.com/journals/uhcm20

# Tweaking the Messages and Approaching the Glass Box: Using AI Chatbots to Promote Help-Seeking for Depressive Symptoms

Jingyuan Shi, KUN Xu & Xiaobei Chen

**To cite this article:** Jingyuan Shi, KUN Xu & Xiaobei Chen (18 Oct 2025): Tweaking the Messages and Approaching the Glass Box: Using AI Chatbots to Promote Help-Seeking for Depressive Symptoms, Journal of Health Communication, DOI: 10.1080/10810730.2025.2569563

**To link to this article:** https://doi.org/10.1080/10810730.2025.2569563



Published online: 18 Oct 2025.

*Journal of Health Communication*, 00: 1–10, 2025

ISSN: 1081-0730 print/1087-0415 online
DOI: https://doi.org/10.1080/10810730.2025.2569563



# Tweaking the Messages and Approaching the Glass Box: Using AI Chatbots to Promote Help-Seeking for Depressive Symptoms

JINGYUAN SHI [1], KUN XU [2], and XIAOBEI CHEN[2]

[1]*Department of Interactive Media, Hong Kong Baptist University, Hong Kong SAR, China*

[2]*College of Journalism and Communications, University of Florida, Gainesville, Florida, USA*

This study tests how AI-based health chatbots' message framing, along with the explanations about human knowledge involvement in their algorithms, influence users' attitudes toward chatbots' recommendation. Based on a two-level human-machine communication framework, an online experiment ($N = 374$) revealed that a chatbot's explanations of the high (vs. low) involvement of human knowledge in its algorithms increased users' trust in the chatbot, which further improved their attitudes toward help-seeking. The message target (targeted vs. mistargeted) employed in the chatbot's recommendations, the involvement of human knowledge in the algorithms, and users' depression tendency jointly influenced users' psychological reactance, which further affected their attitudes toward seeking help from friends and family members. Our findings can contribute to current understandings of how AI shapes the persuasive mechanism of health promotion messages and offer insights into using AI for mental health promotion.

**KEYWORDS**

human-AI communication, persuasive messages, explainable AI, chatbot, human in the loop, depression

Depression is a pressing, global public health condition with increasing prevalence that requires urgent attention (World Health Organization, 2023). In recent years, when scholars have explored the use of artificial intelligence (AI) in mental health care, they have not only emphasized its potential for promoting mental well-being but also recognized the need for more research to improve and refine its performance (Graham et al., 2019). Joining this trend of research, the current study aims to examine the use of AI-based chatbots for promoting individuals' attitudes toward seeking help for depressive symptoms.

Despite the acceleration and expansion of adopting AI technology for health promotion purposes, little is known about what makes AI technology persuasive and how the intended persuasive outcomes can be accomplished. More importantly, the rapid development and adoption of AI leaves individuals with little time to comprehend AI's working mechanisms. Therefore, to understand how individuals react to the persuasive messages and the internal architecture of AI, communication scholars have recently combined literature on human-machine communication (HMC) and explainable AI (XAI) and proposed a two-level HMC framework to address the question (Xu & Shi, 2024). According to this newly developed framework, the first level of HMC centers on how humans directly respond to AI interfaces. The second level underscores the effects of offering explanations of AI's working mechanisms to users and seeks to understand how users perceive, evaluate, and respond to those explanations. Furthermore, this framework proposes that users' evaluation of AI interfaces may interact with their understanding of the black box of AI in determining their attitudes toward and acceptance of AI.

Based on this theoretical framework, we focused on the framing of an AI-based chatbot's health promotion messages (i.e., a first-level HMC variable) and the explanations of human knowledge involvement in the chatbot's algorithms (i.e., a second-level HMC variable). We aim at understanding how the first-level HMC variable and the second-level variable independently and jointly affect users' acceptance of AI's recommendation on promoting mental wellbeing. The findings are expected to advance the theoretical discussion about the persuasive mechanisms of human-AI communication and offer practical insights into using AI for mental health promotion.

Address correspondence to Jingyuan Shi Department of Interactive Media, Hong Kong Baptist University, 5 Hereford Road, Kowloon Tong, Kowloon, Hong Kong SAR, China. E-mail: jolieshi@hkbu.edu.hk

## Message Framing in Human-AI Communication

Although communication scholars have acknowledged that AI complicates the processes of persuasion and their outcomes (Dehnert & Mongeau, 2022; Zarouali et al., 2022), research has rarely explicated the persuasive mechanisms in human–AI communication (Xu & Shi, 2024). Moreover, despite research on how technology interface cues have been used for persuasion (e.g., Oinas-Kukkonen & Harjumaa, 2008), the effects of persuasive messages delivered by technological agents have not been adequately investigated. Thus, the current study aims to investigate how the message framing of persuasive messages delivered by an AI-based chatbot affects users' attitudes toward the AI's recommendations.

### *Targeted versus Mistargeted Messages*

Encouraging help-seeking for depression has long been documented as a challenging task, partly due to the preexisting, persistent negative attitudes toward help-seeking for depression among adults worldwide (Mackenzie et al., 2014). To reduce those negative responses, communication scholars have investigated the role of persuasive messages in promoting help-seeking for depression by assessing the effects of various message options. Based on their results, using mistargeted (vs. targeted) messages has been identified as a promising strategy (Siegel et al., 2015).

With reference to Siegel et al.'s (2015) conceptualizations, we define a persuasive message as being *targeted* when the recipient perceives that their own beliefs, attitudes, and/or behaviors are the target of persuasion. By contrast, we define a persuasive message as being *mistargeted* when the recipient perceives that someone else's beliefs, attitudes, or behaviors, not their own, are the target of persuasion. Research on mistargeted versus targeted messages can be traced to Walster and Festinger's (1962) study. In their experiment, participants overheard a conversation about an issue either highly relevant or highly irrelevant to themselves, and the conversation advocated a position on the issue which was expected to be supported by the involved participants. Walster and Festinger (1962) found that overhearing a conversation influenced highly involved listeners' attitudes, as the listeners believed that discussants were not aware of their existence. The results of subsequent research (e.g., Brock & Becker, 1965) confirmed Walster and Festinger's (1962) findings in which listeners' exposure to the messages that are not seen as being intentionally designed for them may shape their attitudes.

Since then, scholars have focused on the role of mistargeted (vs. targeted) messages in promoting health behaviors. For instance, in a study on anti-inhalant messages, Crano et al. (2007) revealed that, among adolescent audiences, messages that asked parents to help their adolescent children avoid using inhalants were evaluated more favorably than messages that directly urged adolescents to avoid inhalants. In addition, in Siegel et al.'s (2015) experiment, among participants who experienced depressive symptoms, exposure to a printed public service announcement (PSA) using mistargeted messages about depression was more likely to increase their intention to seek help from a close friend and a romantic partner than exposure to one using targeted messages. Although Siegel et al. (2015) did not assess the explanatory mechanism behind their findings, they suggested that the mistargeted messages might have induced less psychological reactance among participants with depression and thus outperformed the targeted messages in promoting help-seeking.

Biased information processing among people with depression could explain why mistargeted messages may outperform targeted ones in promoting help-seeking. According to the cognitive theory of depression (Beck, 1987), individuals with depression tend to have negative views about themselves, the future, and the world. McIntosh and Fischer (2000) have encapsulated those negative views into a single concept, *self-relevant negative attitude*, which has been recognized as a reason for reactance to messages promoting mental health (Siegel et al., 2017). Furthermore, as pinpointed in Wisco's (2009) review article, compared with externally focused content, self-relevant content has exacerbated negative bias among depressed individuals. Thus, it is reasonable to expect mistargeted messages, which are perceived as being less self-relevant than targeted messages, to be processed less negatively among individuals with depressive symptoms and thus induce less psychological reactance. Thus, we propose Hypothesis 1 (H1) and Hypothesis 2 (H2):

> **H1:** An AI-based chatbot's mistargeted messages will induce less psychological reactance than targeted messages among participants with depressive symptoms; however, among participants without depressive symptoms, no difference in inducing psychological reactance will arise between mistargeted and targeted messages.

> **H2:** Psychological reactance will be negatively associated with participants' attitudes toward the chatbot's recommendation on seeking help for depressive symptoms from friends and family members.

## XAI in Human-AI Communication

XAI has been referred to as "the class of systems that provide visibility into how an AI system makes decisions and predictions and executes its actions" (Rai, 2020, p. 138). Recent research has demonstrated the value of XAI in enhancing users' trust in AI systems. For example, research has indicated that users' trust in autonomous agents decreased when the agents' decision-making was not made transparent (Weitz et al., 2019). By contrast, providing reasonings for a conversational agent's recommendation can enhance users' trust in the technology (Wilkinson et al., 2021). Similarly, research suggested that a physically embodied robot's explanations of its own behavior can substantially elevate its perceived trustworthiness (Fischer et al., 2018).

While explanations of AI can increase users' trust, such trust may further bring about users' attitudinal and behavioral changes in their interactions with AI. Along those lines, emerging research has suggested that establishing trust is a key step

in users' adoption of AI systems, as proper justifications of AI's decision-making can amplify users' perceived transparency of and further engagement with such systems (Wilkinson et al., 2021). Nevertheless, what remains to be explored is how different explanations of AI systems influence users' trust in AI. Hence, we introduce the human-in-the-loop approach to explore the effects of different explanations of AI systems.

### *Human-in-the-Loop Approach*

Referred to as the human-in-the-loop approach (Deng et al., 2020; Wu et al., 2022; Zanzotto, 2019), research has suggested that designing algorithms informed by human knowledge can reduce the amount of required data, increase the accuracy of predictions, and combine human experience with machine agency (Deng et al., 2020). Adopting the human-in-the-loop approach in XAI can also give appropriate credit to knowledge producers and elucidate whose knowledge has been included in specific deployments of algorithms (Zanzotto, 2019). For example, in biomedicine and clinical medicine, the inclusion of doctors' knowledge and expertise in algorithms can enhance the reliability of diagnoses and save the time needed for obtaining deployable clinical decisions (Holzinger, 2016; Wu et al., 2022). Thus, the current study uses the human-in-the-loop approach and explores how adding professionals' authoritative health-related knowledge in the AI-based chatbots' algorithms can increase the interpretability of the AI systems and influence users' attitudes toward AI's recommendations.

Although keeping human in the loop of AI's decision-making is pivotal (Adadi & Berrada, 2018), users' responses toward the approach may not always be positive, for the AI systems using this approach may recommend biased results due to human stereotypes and/or biases. Here, the effects of algorithms designed with human knowledge may depend on whether users perceive the tasks of the AI systems as more human-oriented or more machine-oriented. For example, research on machine heuristic has suggested that when it comes to processing credit card information, participants had more trust in an AI agent than a human agent (Sundar & Kim, 2019). Moreover, for tasks requiring mechanical skills (e.g., work scheduling), participants trusted machine agency and human agency equally; nevertheless, for tasks requiring human judgment (e.g., performance evaluations), participants perceived machine decisions as less objective and trustworthy than human decisions (Lee, 2018). Thus, whether and to what extent individuals favor the human-in-the-loop approach may depend on the nature of the AI tasks.

The uncertainty about users' preferences for human involvement in AI-made decisions might be more prominent in the current context of promoting help-seeking for depressive symptoms. On the one hand, according to machine heuristic, individuals may prefer to rely on chatbots' machine nature to protect their privacy, trust machines' objective recommendations, and favor unbiased and non-stereotypical judgments (Aktan et al., 2022; Sundar, 2020). On the other hand, individuals may prefer to receive human-based decisions, for the tasks of understanding users' emotional distress and offering mental health-related recommendations can be regarded as personal and human ones (Aktan et al., 2022). Considering that it remains unknown whether users prefer a machine-based decision that is free from human biases or a machine-based decision that involves human knowledge, we propose the following research question:

**RQ1:** How will an AI-based chatbot's algorithm with high involvement of human knowledge differ from one with low involvement of human knowledge in affecting users' trust in the chatbot?

As reviewed above, prior research has presented trust as a mediator between AI's transparency and users' acceptance of AI's recommendations. Thus, we propose the following research question:

**RQ2:** How will trust mediate the relationship between the different levels of human knowledge involvement in an AI-based chatbot's algorithm and users' attitudes toward the chatbot's recommendation on seeking help for depressive symptoms from friends and family members?

Furthermore, as suggested by the two-level HMC framework, individuals' perceptions and evaluations of AI's features and interface can interact with their understanding of AI's working mechanisms (Xu & Shi, 2024). Additionally, considering that depressed individuals tend to experience cognitive bias, which affects their information processing (Wisco, 2009), it is reasonable to explore the joint effects of a chatbot's message target, individuals' depression tendency, and the chatbot's explanations of human knowledge involvement in its algorithms. To that end, we propose the following research question:

**RQ3:** How will the two-way interaction between message target and depression tendency, proposed in H1, further interact with the different levels of human knowledge involvement in an AI-based chatbots' algorithm in affecting participants' psychological reactance?

## Method

### *Participants*

We recruited 534 adults living in the United States from Qualtrics online panels with age as the primary quota control variable to ensure that the distribution of age was similar to the U.S. population. To prevent any potential harm, we recruited only individuals who were not clinically diagnosed with depression nor had sought professional help for depression in the past year. After applying two attention check questions to ensure data quality, the final sample size included 374 participants, which met the criteria for desired sample size based on power analysis.

Of the 374 participants, 153 were men (40.9%), 281 were women (58.3%), and 3 of them (0.8%) chose "other" as their gender. By age, they were 45.23 years old on average ($SD$ =

15.92). By ethnicity, most were Caucasian (74.1%), followed by Black or African (9.4%), Asian (7.8%), Hispanic or Latino (4.3%), mixed race (3.2%), Native American (0.8%), and other (0.5%).

### *Experimental Stimuli and Design*

Juji, a cognitive AI conversational agent (Xiao et al., 2020; Zhou et al., 2019) was employed here to interact with participants. We built the main chat flow, customized the agent's avatar, and controlled the conversation topics to fit our research purpose. Our experiment followed a 2 (message framing: mistargeted vs. targeted) × 2 (human knowledge involvement: high vs. low) between-subjects factorial design.

To manipulate the message targets, mistargeted messages were designed to prompt participants to feel that the chatbot was discussing depressive symptoms that could occur to other people, whereas targeted messages were designed to prompt participants to feel that the chatbot was discussing depressive symptoms that they themselves could experience. We designed three messages for each condition, and all of them presented basic knowledge and facts about depression and explicitly encouraged seeking help for depressive symptoms.

In the high human knowledge involvement conditions, participants were informed that the chatbot was pre-installed with health-related information provided by health professionals. In the low human knowledge involvement conditions, participants were informed that the chatbot was installed with a group of machine learning techniques, which allowed it to collect online health-related information by itself. The stimuli messages and screenshots of the chatbot's interface were presented in Appendices.

### *Experimental Procedures*

After indicating their consent, participants first answered questions about demographics and depression tendency. Next, they were told that they would interact with a prototype information-driven chatbot. The chatbot was programmed into the Qualtrics online questionnaire using HTML and JavaScript codes, which allowed participants to complete their interaction with the chatbot without having to switch between different browser tabs.

Participants were then exposed to a chatbot with the gender-neutral name "Sai." Participants were told by the chatbot that it was developed to provide information about mental well-being. They were asked by the chatbot to make a brief self-introduction at the beginning of the conversation. After responding to participants' introduction, the chatbot asked them whether they had ever searched for health-related information online. Whether participants answered yes or no, the conversation was pre-programmed to smoothly transition to the topic of depression-related information. These steps were designed as an ice-breaking approach to familiarize participants with the chatbot's conversation style and response speed.

Next, the chatbot provided a message about what algorithms it used to provide information. After that, the conversation flow led the chatbot to provide three depression-related messages to the participants. After each message, participants were asked by the chatbot to rate the readability of the messages using a close-ended question. This step was designed to (1) allow participants to continue interacting with the chatbot, (2) prevent the chatbot from continuously delivering messages without pausing, (3) increase the variance of the types of human–chatbot interaction, and (4) minimize the effects of participants' open-ended responses as a potential confounding variable. After delivering all the messages, the chatbot thanked the participants, and the system directed them back to the remaining questionnaire items.

### *Measures*

#### *Trust*

Trust was measured by adapting a measure of perceived trustworthiness used in past research on human–computer interaction (Gong & Nass, 2007). Participants rated how well each adjective described the chatbot on a 10-point semantic differential scale with five items: (a) "Untrustworthy" to "Trustworthy," (b) "Unreliable" to "Reliable," (c) "Inconsiderate" to "Considerate," (d) "Dangerous" to "Safe," and (e) "Dishonest" to "Honest," $\alpha = .94$., $M = 8.11$, $SD = 1.72$.

#### *Psychological Reactance*

Using measures from Dillard et al. (2023), we employed six items to assess reactance in two dimensions: anger and negative cognition. Participants indicated to what extent they felt (a) "Angry," (b) "Annoyed," (c) "Irritated," and (d) "Aggravated" while viewing the messages provided by the chatbot on a 7-point scale ranging from 1 (*none of this feeling*) to 7 (*a great deal of this feeling*). Participants also reported to what extent they agreed with two statements on a 7-point scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*): (a) "I have critical thoughts about the messages from the chatbot," and (b) "I think that I dislike the messages from the chatbot." We averaged the responses to the six items on an index of reactance ($\alpha = .92$, $M = 2.12$, $SD = 1.30$).

#### *Attitudes Toward AI's Recommendation*

Participants rated their attitudes toward seeking help from their friends and family members when they experience depressive symptoms (i.e., the chatbot's recommendation) using three 7-point semantic differential items: (a) "Unimportant" to "Important," (b) "Worthless" to "Valuable," and (c) "Undesirable" to "Desirable," $\alpha = .96$, $M = 5.12$, $SD = 1.76$.

#### *Depression Tendency*

We assessed participants' depression tendency using a seven-item depression scale drawing from the short Depression Anxiety Stress Scale (DASS; Lovibond & Lovibond, 1995). Participants rated to what extent each item applied to them during the past week on a four-point scale including 0 (*Did not apply to me at all*), 2 (*Applied to me to some degree*) 4 (*Applied to me to a considerable degree*), and 6 (*Applied to me very much*). Following the DASS's instructions, we totaled the points given on seven items ($\alpha = .96$, range = 7–28 points), such that a total score greater than 13 indicated a need for further counseling and clinical examination for depression (Lovibond & Lovibond, 1995). We then recoded the variable into a new

binary variable labeled "Depressive tendency," for which participants who scored over 13 ($n = 104$, 27.8%) were coded as 1 (*yes*).

*Covariates*

We controlled gender (1 = *man*, 0 = *woman*, Missing = *other*) in all the analyses because it was found to influence individuals' attitudes toward AI (e.g., Xu, 2019) and attitudes toward seeking help for depression (Mackenzie et al., 2014).

### *Data Analysis*

H1 and RQ3 examined how experimental manipulations and participants' depression tendency influenced psychological reactance. We conducted an analysis of covariance (ANCOVA) with psychological reactance being the dependent variable. We then used PROCESS Macro models to test the mediation effects of reactance on attitudes as proposed by H2. For RQ1 and RQ2, which explored the effects of experimental manipulations on trust and its mediation effect on attitudes, we conducted an ANCOVA with trust being the dependent variable and tested the mediation effects using PROCESS Macro models.

## Results

### *Manipulation Check*

Participants rated the chatbot's depression-related messages with two items using a 7-point bipolar scale: (a) "The chatbot was generally talking about a situation that could occur to me/to others" and (b) "The chatbot asked me whether I have ever/I know someone who has experienced depressive symptoms" (Spearman-Brown $r = .59$, $p < .001$, $\alpha = .74$, $M = 3.76$, $SD = 2.09$). A higher score indicated participants' greater belief that the messages were about someone else (i.e., mistargeted messages). Independent T-test revealed that participants in the mistargeted message conditions ($M = 5.21$, $SD = 1.63$) rated significantly higher on the scale than their counterparts in the targeted conditions ($M = 2.27$, $SD = 1.32$), $t(360.34) = 19.14$, $p < .001$, Cohen's $d = 1.97$.

Participants also reported to what extent they agreed that the chatbot searched, collected, and analyzed the information by itself (1 = *strongly disagree*, 7 = *strongly agree*). Independent T-test suggested that those assigned to the low human knowledge conditions ($M = 5.43$, $SD = 1.42$) scored significantly higher on the item compared to those assigned to the high human knowledge conditions ($M = 4.10$, $SD = 1.82$), $t(349.00) = 7.83$, $p < .001$, Cohen's $d = 0.81$. Thus, the experimental manipulation was successful.

### *Major Findings*

To test H1 and answer RQ3, we conducted an ANCOVA with experimental manipulations and participants' depression tendency being independent variables, psychological reactance being treated as a dependent variable, and gender being a covariate. Results revealed a significant three-way interaction, $F(1, 362) = 8.55$, $p = .004$, $\eta^2 = .022$. In detail, mistargeted messages induced significantly less reactance ($M = 1.72$, $SD = 0.86$) than targeted ones ($M = 2.16$, $SD = 1.10$) among participants without depressive symptomology when the chatbot presented explanations of low human knowledge involvement in its algorithms, $p = .034$. Thus, the data were inconsistent with H1.

H2 predicted that reactance would be negatively associated with attitudes toward the chatbot's recommendations. Given that a three-way interaction on reactance was observed, we conducted a *moderated moderated mediation* analysis using PROCESS Macro Model 11 with 5,000 bootstrapped samples (Hayes, 2018). The analysis revealed a significant three-way interaction on reactance, $b = 1.60$, $t = 2.92$, $p = .004$, which in turn was negatively associated with participants' attitudes toward the chatbot's recommendations, $b = -0.20$, $t = -2.57$, $p = .010$. As shown in Figure 1, for participants without depression tendency, only when the chatbot provided explanations of low human knowledge involvement in its algorithms did participants' psychological reactance mediate the effect of message target on their attitudes toward help-seeking from family members and friends, $b = -0.09$, 95% CI [−0.20, −0.007], index of moderated moderated mediation = −0.32, 95% CI [−0.74, −0.03]. Therefore, the data were consistent with H2.

To answer RQ1, we conducted an ANCOVA with experimental manipulations as independent variables, trust as dependent variable, and participants' depressive tendency as well as gender as covariates. Results suggested that the chatbot's high human knowledge involvement algorithm ($M = 8.33$, $SD = 1.68$) elicited users' greater trust in the chatbot than its low human knowledge involvement algorithm ($M = 7.97$, $SD = 1.64$), $F(1, 365) = 4.92$, $p = .027$, $\eta^2 = .013$. We also conducted mediation analyses using PROCESS Macro Model 4 with 5,000

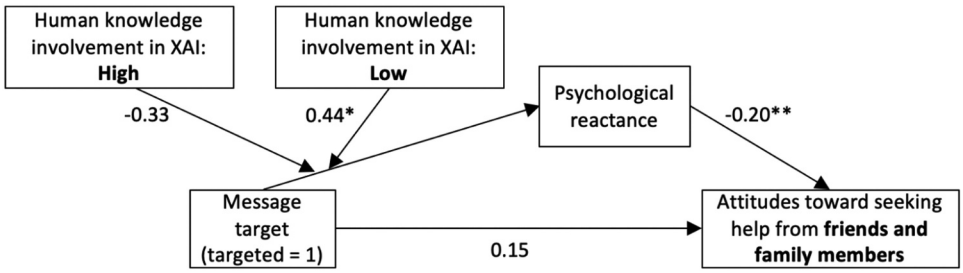


**Figure 1.** Path coefficients of moderated moderated mediation analyses on attitudes toward help-seeking among participants without depression tendency. *Note*. Path coefficients are unstandardized. Gender was controlled in the models. *$p < .05$. **$p < .01$. ***$p < .001$.

**Table 1.** Indirect, direct, and total effects of different levels of human knowledge involvement in XAI on users' attitudes through trust toward the chatbot

| | | 95% Confidence Interval | |
|---|---|---|---|
| | *b* (SE) | LLCI | ULCI |
| Attitudes toward seeking help from friends and family members | | | |
| Indirect effect | 0.10 (.05) | 0.02 | 0.22 |
| Direct effect | 0.16 (.17) | −0.17 | 0.49 |
| Total effect | 0.27 (.17) | −0.07 | 0.61 |

Note. *b:* unstandardized coefficient. SE: standard error. LLCI: lower level confidence interval. ULCI: upper level confidence interval.

bootstrapped samples (Hayes, 2018) to answer RQ2. Results suggested that the indirect effect of the explanations of high human knowledge involvement on users' attitudes through trust was significant, $b = 0.10$, 95% CI [0.02, 0.22] (see Table 1).

## Discussion

Drawing on the two-level HMC framework, the current study tested an AI-based chatbot that employed persuasive messages and presented its algorithms with variations in human knowledge involvement. Our findings discovered two routes that encouraged users' acceptance of AI's recommendation: 1) enhancing users' trust in AI and 2) reducing users' psychological reactance toward AI's persuasive messages. In that light, this study is one of the first to understand users' psychological processing of AI's persuasion, and its findings can make theoretical contributions to human-AI communication as well as offer practical insights into user interface design and mental health promotions.

### *Main Findings and Theoretical Implications*

This study first discovered that explanations about high (vs. low) involvement of human knowledge in a chatbot's algorithms evoked greater trust in the chatbot, which further improved users' attitudes toward the chatbot's recommendation on help-seeking. The findings not only corroborated the previous research findings that explanations and transparency of AI systems can influence users' trust (Liu, 2021; Wilkinson et al., 2021) but also demonstrated that such trust could lead to persuasive outcomes of human-AI communication in healthcare settings.

The current findings also revealed a conditional effect of AI's persuasion on users' attitudes toward its recommendations. In particular, the effects of the message target (mistargeted vs. targeted) on users' attitudes toward help-seeking for depression via psychological reactance were contingent on individuals' depression tendency and the extent of human knowledge involvement in AI's algorithms.

Considering that our results about low human knowledge involvement in algorithms undermined users' trust in the chatbot, it is plausible that trust did not function as a mental shortcut for participants to make a quick judgment about the chatbot's recommendations under this condition. Rather, when the chatbot's recommendation algorithms were perceived to have low human knowledge involvement, participants had to scrutinize the content of the messages delivered by the chatbot. This might explain why the involvement of human knowledge in the algorithms was a *contingent* moderator in our results (see Holbert & Park, 2020), such that the message target induced different levels of reactance in low human knowledge involvement conditions (depending on users' depression tendency), but not in high human knowledge involvement conditions.

Participants' depression tendency also emerged as a *contingent* moderator in the findings. One explanation may be that compared with non-depressed participants, participants exhibiting depression tendency reported greater reactance toward the messages delivered by the chatbot, which is consistent with the previous finding that those with depression tendency tend to have strong reactance toward messages promoting mental health (Siegel et al., 2017). It is also possible that although employing mistargeted messages was able to alleviate reactance among non-depressed participants, it was not potent enough to reduce depressed participants' reactance as their reactance may be more persistent.

The statistically significant difference between participants depression tendency in their reactance induced by the message target can also be attributed to the different communication contexts in previous studies versus ours. Specifically, in Crano et al.'s (2007) and Siegel et al.'s (2015) studies, participants were exposed to mass communication, and the highly accessible nature of communication made the messages less self-relevant. Along those lines, targeted versus mistargeted messages could be clearly differentiated based on recipients' perception of their relevance to them. For message recipients with depressive symptoms, they were able to recognize this difference in self-relevance induced by message target and respond less negatively toward the mistargeted ones. Nevertheless, in the current study, messages were delivered to each participant individually by an AI-based chatbot, and there was no other recipient of those particular messages. Such low accessibility can be expected to increase the self-relevance of the messages compared to mass communication messages. For participants with depressive symptoms who tend to negatively process self-relevant information, they may not be able to distinguish the levels of self-relevance induced by mistargeted messages as the communication context itself is private and exclusive. Thus, when the chatbot with low human knowledge involvement in its XAI system delivered mental health promotion messages, message target only made a difference in reducing reactance among participants without depression tendency.

Despite the consistency with previous findings indicating the superiority of mistargeted messages in promoting help-seeking for depressive symptoms (Siegel et al., 2015), our findings documented a different boundary condition from what has been found in the context of exposure to PSA in mass communication (Crano et al., 2007; Siegel et al., 2015). In our study, when delivered by an AI-based chatbot that provided explanations about its low human knowledge involvement in the algorithms, mistargeted messages induced less reactance than targeted messages among non-depressed participants. However, such effect

faded away among depressed participants. The difference between our study findings and the previous ones could be attributed to the different communication contexts.

### *Practical Implications*

The current findings revealed that the effect of AI's health promotion messages is contingent upon users' perceptions and understanding of AI's internal algorithms. Therefore, when employing AI tools in health promotion, health professionals' involvement with the AI systems and the transparency of the systems are critical in shaping users' acceptance of AI's recommendations. In addition, the current findings show that persuasive messages found to be effective in traditional health promotion may not have the identical effects as in human–AI communication. After all, differences in communication contexts and in explanations of AI algorithms can alter how individuals interpret the messages. More formative research is therefore needed to elucidate the effects of message features when planning health promotions involving AI-based technology.

### *Limitations and Future Research*

First, the forms in which human knowledge is kept in the loop can be diverse. Aside from informing participants of the amount of human knowledge involved in chatbots' algorithms, future research could review other ways in which human knowledge is incorporated and framed. Second, the two primary variables in this study are only a small subset of concepts derived from human-AI communication. More research is needed to fully comprehend how human-AI communication could be employed to promote mental wellbeing. Third, the experiment required participants to interact with the chatbot and accessed the persuasive outcomes afterward. Future research could employ field experiments and longitudinal designs to further understand how AI could assist in mental health promotion in real-life contexts.

## Conclusion

Overall, the study findings suggest that AI-based technologies influence users' acceptance of its recommendations through two paths. One path highlights the use of persuasive message strategies with which chatbots can be designed to facilitate attitudinal outcomes, and the other underscores the importance of explaining the role of human intelligence in algorithmic recommendations. Altogether, our work demonstrates that using AI-based technology for mental health promotion calls for research focused on the interactions among individual differences, persuasive messages, and AI explanations.

## Disclosure Statement

No potential conflict of interest was reported by the author(s).

## Funding

The work was supported by the Hong Kong Baptist University [Rising Star Research Grant].

## ORCID

Jingyuan Shi http://orcid.org/0000-0001-5675-4512
KUN Xu http://orcid.org/0000-0001-9044-821X

## References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Aktan, M. E., Turhan, Z., & Dolu, İ. (2022). Attitudes and perspectives towards the preferences for artificial intelligence in psychotherapy. *Computers in Human Behavior*, *133*, Article 107273. https://doi.org/10.1016/j.chb.2022.107273

Beck, A. T. (1987). Cognitive models of depression. *Journal of Cognitive Psychotherapy*, *1*, 5–37.

Brock, T. C., & Becker, L. A. (1965). Ineffectiveness of "overheard" counterpropaganda. *Journal of Personality & Social Psychology*, *2*(5), 654–660. https://doi.org/10.1037/h0022720

Crano, W. D., Siegel, J. T., Alvaro, E. M., & Patel, N. M. (2007). Overcoming adolescents' resistance to anti-inhalant appeals. *Psychology of Addictive Behaviors*, *21*(4), 516–524. https://doi.org/10.1037/0893-164X.21.4.516

Dehnert, M., & Mongeau, P. A. (2022). Persuasion in the age of artificial intelligence (AI): Theories and complications of AI-based persuasion. *Human Communication Research*, *48*(3), 386–403. https://doi.org/10.1093/hcr/hqac006

Deng, C., Ji, X., Rainey, C., Zhang, J., & Lu, W. (2020). Integrating machine learning with human knowledge. *Iscience*, *23*(11), Article 101656. https://doi.org/10.1016/j.isci.2020.101656

Dillard, J. P., Tian, X., Cruz, S. M., Smith, R. A., & Shen, L. (2023). Persuasive messages, social norms, and reactance: A study of masking behavior during a COVID-19 campus health campaign. *Health Communication*, *38*(7), 1338–1348. https://doi.org/10.1080/10410236.2021.2007579

Fischer, K., Weigelin, H. M., & Bodenhagen, L. (2018). Increasing trust in human-robot medical interactions: Effects of transparency and adaptability. *Paladyn, Journal of Behavioral Robotics*, *9*(1), 95–109. https://doi.org/10.1515/pjbr-2018-0007

Gong, L., & Nass, C. (2007). When a talking-face computer agent is half-human and half-humanoid: Human identity and consistency preference. *Human Communication Research*, *33*(2), 163–193. https://doi.org/10.1111/j.1468-2958.2007.00295.x

Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H.-C., & Jeste, D. V. (2019). Artificial intelligence for mental health and mental illnesses: An overview. *Current Psychiatry Reports*, *21*(11), 1–26. https://doi.org/10.1007/s11920-019-1094-0

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (3rd ed.). Guilford Press.

Holbert, R. L., & Park, E. (2020). Conceptualizing, organizing, and positing moderation in communication research. *Communication Theory*, *30*(3), 227–246. https://doi.org/10.1093/ct/qtz006

Holzinger, A. (2016). Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Informatics*, *3*(2), 119–131. https://doi.org/10.1007/S40708-016-0042-6

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1), 1–16. https://doi.org/10.1177/2053951718756684

Liu, B. (2021). In AI we trust? Effects of agency locus and transparency on uncertainty reduction in human-AI interaction. *Journal of Computer-Mediated Communication*, *26*(6), 384–402. https://doi.org/10.1093/jcmc/zmab013

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the depression anxiety stress scales

(DASS) with the Beck depression and anxiety inventories. *Behaviour Research and Therapy*, *33*(3), 335–343. https://doi.org/10.1016/0005-7967(94)00075-U

Mackenzie, C. S., Erickson, J., Deane, F. P., & Wright, M. (2014). Changes in attitudes toward seeking mental health services: A 40-year cross-temporal meta-analysis. *Clinical Psychology Review*, *34*(2), 99–106. https://doi.org/10.1016/j.cpr.2013.12.001

McIntosh, C. N., & Fischer, D. G. (2000). Beck's cognitive triad: One versus three factors. *Canadian Journal of Behavioral Science*, *32*(3), 153–157. https://doi.org/10.1037/h0087110

Oinas-Kukkonen, H., & Harjumaa, M. (2008). A systematic framework for designing and evaluating persuasive systems. In *Proceedings of the Third International Conference on Persuasive Technology* Oulu, Finland (pp. 164–176). https://doi.org/10.1007/978-3-540-68504-3_15

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5

Siegel, J. T., Lienemann, B. A., & Rosenberg, B. D. (2017). Resistance, reactance, and misinterpretation: Highlighting the challenge of persuading people with depression to seek help. *Social and Personality Psychology Compass*, *11*, Article e12322. https://doi.org/10.1111/spc3.12322

Siegel, J. T., Lienemann, B. A., & Tan, C. N. (2015). Influencing help-seeking among people with elevated depressive symptomatology: Mistargeting as a persuasive technique. *Clinical Psychological Science*, *3*(2), 242–255. https://doi.org/10.1177/2167702614542846

Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human-AI interaction (HAII). *Journal of Computer-Mediated Communication*, *25*(1), 74–88. https://doi.org/10.1093/jcmc/zmz026

Sundar, S. S., & Kim, J. (2019, May). Machine heuristic: When we trust computers more than humans with our personal information. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* Glasgow, UK (pp. 1–9). https://doi.org/10.1145/3290605.3300768

Walster, E., & Festinger, L. (1962). The effectiveness of "overheard" persuasive communications. *Journal of Abnormal and Social Psychology*, *65*(6), 395–402. https://doi.org/10.1037/h0041172

Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2019). "Do you trust me?" Increasing user-trust by integrating virtual agents in explainable AI interaction design. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* Paris, France (pp. 7–9). ACM. https://doi.org/10.1145/3308532.3329441

Wilkinson, D., Alkan, Ö., Liao, Q. V., Mattetti, M., Vejsbjerg, I., Knijnenburg, B. P., & Daly, E. (2021). Why or why not? The effect of justification styles on chatbot recommendations. *ACM Transactions on Information Systems*, *39*(4), 1–21. https://doi.org/10.1145/3441715

Wisco, B. E. (2009). Depressive cognition: Self-reference and depth of processing. *Clinical Psychology Review*, *29*(4), 382–392. https://doi.org/10.1016/j.cpr.2009.03.003

World Health Organization. (2023). *Depression*. https://www.who.int/news-room/fact-sheets/detail/depression

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, *135*, 364–381. https://doi.org/10.1016/j.future.2022.05.014

Xiao, Z., Zhou, M. X., Chen, W., Yang, H., & Chi, C. (2020, April). If I hear you correctly: Building and evaluating interview chatbots with active listening skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* Honolulu, Hawai'i, USA (pp. 1–14). https://doi.org/10.1145/3313831.3376131

Xu, K. (2019). First encounter with robot alpha: How individual differences interact with vocal and kinetic cues in users' social responses. *New Media and Society*, *21*(11–12), 2522–2547. https://doi.org/10.1177/1461444819851479

Xu, K., & Shi, J. (2024). Visioning a two-level human-machine communication framework: Initiating conversations between explainable AI and communication. *Communication Theory*, *qtae016*(4), 216–229. https://doi.org/10.1093/ct/qtae016

Zanzotto, F. M. (2019). Human-in-the-loop artificial intelligence. *The Journal of Artificial Intelligence Research*, *64*, 243–252. https://doi.org/10.1613/jair.1.11345

Zarouali, B., Boerman, S. C., Voorveld, H. A., & van Noort, G. (2022). The algorithmic persuasion framework in online communication: Conceptualization and a future research agenda. *Internet Research*, *32*(4), 1076–1096. https://doi.org/10.1108/INTR-01-2021-0049

Zhou, M. X., Chen, W., Xiao, Z., Yang, H., Chi, T., & Williams, R. (2019). Getting virtually personal: Chatbots who actively listen to you and infer your personality. In *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion* Greenville, South Carolina, USA (pp. 123–124). ACM. https://doi.org/10.1145/3308557.3308667

**Appendix A. All Stimuli Messages Delivered by the Chatbot**

A Message About Human Knowledge Involvement

- Low human knowledge involvement
  I was installed with a group of machine learning techniques, which allowed me to collect online health-related information by myself. I search, scrape, and analyze information based on health professionals' online talks, online interview scripts, blogs, and monographs every twice a week. I parse out massive information and seek to provide the most up-to-date and credible recommendations.

- High human knowledge involvement
  I was pre-installed with health-related information provided by health professionals. Health professionals search, collect, and analyze health-related information based on their knowledge and expertise every twice a week. These health professionals parse out the massive information, send it to me, and seek to provide the most up-to-date and credible recommendations.

Targeted Messages with Different Levels of Human KnowledgeInvolvement

- Have you ever experienced depressive symptoms? You may feel sad, hopeless, and lose interest in things you used to enjoy. You may also feel a lack of energy, have difficulties in falling asleep at night, and experience changes in appetite or weight. [More importantly, based on the machine learning techniques that enable me to collect and analyze the information online/ More importantly, based on the information that is collected and analyzed by health professionals], such symptoms persist for weeks or months and are bad enough to interfere with your work, social life, and family life.

- [According to health professionals' online talks, interviews, and blogs, which I used as part of the machine learning process/According to the information provided by health professionals, which was pre-programmed into my server], experiencing those depressive symptoms, however, is not your fault, and you are not weak or worthless. Having a mental issue does not mean you are "crazy." What it does mean is that there is an illness, like a flu, that is affecting your brain and that you are likely less in control of how you want to live your life.

- [Based on the massive information I analyzed using machine learning/Based on the massive information analyzed and provided by health professionals], here are some recommendations. If you are experiencing depressive symptoms, please talk to people around you. Friends and family are often integral in noticing and addressing problematic symptoms, and can be a major part of the recovery process. It's not usual for someone to disclose their personal challenges and distress. However, a caring friend or family member will put you at ease. They are the people who will listen to you and help you find support.

Mistargeted Messages with Different Levels of Human KnowledgeInvolvement

- Do you know someone who has experienced depressive symptoms? They may feel sad, hopeless, and lose interest in things they used to enjoy. They may also feel a lack of energy, have difficulties in falling asleep at night, and experience changes in appetite or weight. [More importantly, based on the machine learning techniques that enable me to collect and analyze the information online/More importantly, based on the information that is collected and analyzed by health professionals], such symptoms persist for weeks or months and are bad enough to interfere with their work, social life, and family life.

- [According to health professionals' online talks, interviews, and blogs, which I used as part of the machine learning process/According to the information provided by health professionals, which was pre-programmed into my server], experiencing those depressive symptoms, is not their fault, and they are not weak or worthless. Having a mental issue does not mean someone is "crazy." What it does mean is that there is an illness, like flu, that is affecting someone's brain and the person is less in control of how he/she wants to live his/her life.

- [Based on the massive information I analyzed using machine learning/ Based on the massive information analyzed and provided by health professionals], here are some recommendations. If you know someone who is experiencing depressive symptoms, please talk to them. Friends and family are often integral in noticing and addressing problematic symptoms and can be a major part of the recovery process. It's not usual for someone to disclose their personal challenges and distress. However, your caring will put them at ease. Listening to them and helping them find support are the things you can do.

## Appendix B. Screenshots of Chatbot Interaction Interface