Who Hears My Secrets? Self-Disclosure in Text, Voice, and Robotic AI Counseling:

How Context Collapse, Contextual Integrity Explain Cross-Platform Differences

Jihye Lee
AIX School
Gwangju Institute of Science and Technology
Gwangju, South Korea
cnlgid_@naver.com

ABSTRACT

As artificial intelligence increasingly enters therapeutic roles, the question arises: will people truly open up to chatbots, voice assistants, or social robots? This study examines self-disclosure in AI-based counseling by comparing text-based chatbots, voice-based AI speakers, and physical social robots. We plan to recruit 90 to 120 adults and randomly assign them to one of these three media. Each participant will receive identical scripted prompts regarding recent stressors or personal concerns, and self-disclosure will be measured through both objective session transcripts and subjective post-survey assessments. We also explore how perceptions of context collapse, contextual integrity, and trust differ across platforms and influence disclosure. We hypothesize that text-based chatbots may foster freer expression due to reduced concerns about eavesdropping, whereas voice-based AI may raise worries about unintended audiences. Physical robots, by contrast, could either enhance a sense of presence conducive to sharing or exacerbate surveillance fears linked to cameras and sensors. By including pre-survey measures of privacy orientation and counseling motivation, we account for individual differences in disclosure patterns. Our findings aim to inform best practices for designing AI-assisted psychotherapy, ensuring a balance between privacy, user confidence, and therapeutic efficacy.

CCS CONCEPTS

• Human-centered computing → User studies • Applied computing → Health informatics

KEYWORDS

AI-Based Psychotherapy, Self-Disclosure, Context Collapse, Contextual Integrity

ACM Reference format:
Jihye Lee. 2025. Who Hears My Secrets? Self-Disclosure in Text, Voice, and Robotic AI Counseling: How Context Collapse, Contextual Integrity Explain Cross-Platform Differences. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI'25). ACM, Yokohama, Japan, 3 pages.*

CHI '25 Workshop on Envisioning the Future of Interactive Health, Yokohama, Japan

1 Introduction

Recently, **AI-based psychotherapy** has expanded into various media, including text-based chatbots (e.g., ChatGPT), voice-based AI speakers, and physical robots (social robots), offering accessibility and convenience to individuals with psychological concerns [1, 2]. For example, there have been a series of reports on VR-based AI counseling attempts [3], studies using conversational AI for cognitive bias correction [4], using voice assistants for counseling [5], applying ChatGPT for psychotherapy [6], and implementing social robots in psychiatric care [7]. However, how effectively AI counseling elicits **self-disclosure** from actual users remains insufficiently examined.

Self-disclosure is central to psychotherapy, as clients must candidly reveal their personal experiences, emotions, and concerns for emotional relief, therapeutic alliance, and motivational enhancement to take effect [8, 9]. Yet when users disclose information to an AI medium, multiple factors come into play. Previous literature indicates that individual traits (e.g., privacy orientation, motivation, personality), systemic factors (social presence, technological transparency, trust), and contextual elements (interaction purpose, potential audience) all influence self-disclosure [10, 11]. In particular, this study focuses on two major concepts—"context collapse" and "contextual integrity"—to investigate the external

anxieties (the range of audience, how one's information might be utilized) that users experience in AI counseling [12, 13].

Moreover, trust emerges as a significant variable. When context collapse and contextual integrity issues are heightened, users may regard the system as failing to safeguard their information, thereby lowering trust and curbing self-disclosure [14, 15]. While some have suggested a research design treating trust as a mediator, there is little empirical work clarifying exactly how mediation occurs. In the current study, we measure trust in a post-session survey; if data show that elevated context collapse/contextual integrity concerns → (trust ↓) → self-disclosure ↓, we will interpret this as a potential mediating effect.

Accordingly, our objective is to determine which medium—text (chatbot), voice (AI speaker), or physical robot (social robot)— best facilitates self-disclosure in a psychotherapeutic context, and to explore how contextual factors (context collapse, contextual integrity) and trust account for these differences. This investigation will suggest ways to alleviate user concerns about "who might overhear me?" or "could my information be repurposed?" and to foster trust so as to promote self-disclosure in AI counseling environments.

2. Research Design & Methods

2.1 Research Questions

1. RQ1: In a psychotherapeutic setting, how do text-, voice-, and robot-based AI platforms differ in terms of self-disclosure (both behavioral and subjective measures)?
2. RQ2: How do context collapse, contextual integrity, and trust explain such cross-platform differences in self-disclosure?

If necessary, user-specific factors, such as privacy orientation and counseling motivation, will be tested as moderators.

2.2 Participants & Procedure

This study will recruit 90 to 120 adults aged 20 or older, randomly assigning them to one of three groups based on the AI counseling medium (text, voice, or robot). Before participating in the main counseling session, all individuals will complete a brief pre-survey. This survey will include mandatory demographic questions (e.g., age, gender, and other relevant background information) and, if deemed necessary, a set of short scales (3 to 5 items each) measuring privacy orientation and counseling motivation.

Following this initial survey, participants will engage in a media-based counseling session lasting approximately 10 to 15 minutes. A standardized script will prompt them to discuss typical counseling topics such as recent stressors or concerns, ensuring that all groups receive similar content. In the text-based condition, participants will interact with a chatbot (for example, ChatGPT or a similar system). In the voice-based condition, they will converse with an AI speaker prototype that incorporates speech recognition and basic counseling dialogue. In the robot-based condition, they will communicate with a social robot capable of speech synthesis and expressive behaviors, using the same counseling script employed in the other conditions.

Upon completing this counseling session, participants will fill out a post-survey designed to capture both subjective self-disclosure and their perceptions of the counseling medium. They will be asked, for instance, how candidly they felt they shared personal topics during the session (self-disclosure), whether they believed unintended individuals might have access to their conversation (context collapse), whether they worried about their personal information being used outside of the intended therapeutic context (contextual integrity), and whether they considered the AI medium secure in handling their personal data (trust).

2.3 Measurement & Analysis

In terms of measuring self-disclosure—the primary dependent variable—both objective and subjective approaches will be employed. First, the session logs (audio transcripts for voice/robot conditions; chat data for the text condition) will be collected to quantify aspects such as word count and occurrences of sensitive or emotionally significant statements (e.g., references to family issues, expressions of distress). Second, participants' subjective perceptions of how openly they disclosed personal information will be drawn from their post-survey responses.

For data analysis, we will conduct an analysis of variance (ANOVA) to determine whether self-disclosure differs significantly among the three AI media (text, voice, robot), thereby addressing RQ1. In addition, regression or correlation analyses will explore how context collapse, contextual integrity, and trust relate to the levels of self-disclosure, in line with RQ2. If necessary, privacy orientation and counseling motivation—collected during the pre-survey—may be introduced into the analytic models as moderators or control variables, allowing a more nuanced understanding of how these individual differences influence the main effects of the AI medium on self-disclosure.

3. Expected Findings

The expected results can be summarized in three primary points. First, users' self-disclosure levels will vary across text (chatbot), voice (AI speaker), and physical robot (social robot). Text-based AI, offering anonymity and non-face-to-face interaction, likely involves less context collapse anxiety, thus potentially permitting freer expression of emotions and sensitive content—leading to higher self-disclosure.

On the other hand, voice-based AI speakers are deployed in real environments such as at home or in the workplace, making it easy for users to imagine unintended audiences overhearing them. For instance, family members or colleagues might accidentally listen in, or the device may be continuously recording—fueling stronger awareness of context collapse. Consequently, self-disclosure may be lower in a voice-based scenario.

Meanwhile, physical robots (social robots), by virtue of their humanlike presence (e.g., facial expressions, gestures), can foster rapport that facilitates self-disclosure. However, their built-in camera or sensors might be perceived as always monitoring or recording, resulting in surveillance-related anxiety that curbs information sharing. Consequently, robots might display "two-sided" effects, neither eliciting extremely high disclosures nor extremely low ones, or producing widely varied responses depending on individual user perceptions.

Second, these cross-platform differences are likely intertwined with context collapse, contextual integrity, and trust. In the voice condition, "someone else in my household or office can overhear" represents a clear context collapse worry, which might substantially limit self-disclosure. In the robot group, participants could fear the robot is capturing them on camera or storing audio. By contrast, text-based chatbots do not involve local eavesdroppers but may trigger concerns about "server-based data storage" and potential violation of contextual integrity if user data are used for advertising or big data analytics. Trust may mediate or moderate these concerns; if participants feel "this system is reliable," they might continue disclosing despite worries, whereas low trust could amplify small anxieties into significant disclosure blocks.

Third, individual factors like privacy orientation and counseling motivation might further refine these media effects. Generally, those with high privacy orientation could withhold disclosure regardless of medium, yet they may find text-based AI relatively safer (no local overhearing). Those with high counseling motivation might persist in disclosing sensitive details for the sake of therapeutic gains, even if they harbor moderate anxieties about context collapse or data usage. Thus, user-level differences can shape more nuanced outcomes regarding each medium's ability to elicit self-disclosure.

In sum, text-based chatbots appear most likely to produce robust self-disclosure, whereas voice-based AI speakers may prompt lower disclosure. Social robots, however, may yield more varied outcomes. This is presumably explained by the interplay of context collapse, contextual integrity, and trust, as well as moderated by individual privacy concerns or motivation for therapy.

4. Discussion & Conclusion

This study compared self-disclosure across three media—text-based chatbots, voice-based AI speakers, and physical social robots—and examined the findings through the lenses of context collapse, contextual integrity, and trust. From an "Understanding Users" perspective, our approach highlights concrete anxieties and conveniences that people experience in real therapeutic contexts, thus suggesting user requirements and design implications for HCI and software engineering. When developing or refining AI-based counseling systems, we also draw attention to groups that may feel excluded or uncomfortable—for instance, individuals living with others who worry about being overheard, or older adults who find new devices intimidating. While text-based chatbots may encourage greater self-disclosure by mitigating concerns about context collapse, the physical presence of robots can be both comforting and unsettling, due to the potential for surveillance or recording. These results underscore the importance of offering personalized privacy settings and conducting usability tests tailored to different types of users.

Moreover, by quantitatively and qualitatively analyzing differences in self-disclosure, we establish an initial evidence base for how people with strong counseling motivation or a high privacy orientation may respond differently to each medium. Such insights can directly inform requirements analysis in HCI design and software engineering, helping to determine which medium is truly "replicable," "inclusive," and capable of delivering "practical significance" in psychotherapy settings. In future work, systematically manipulating design elements (e.g., automated volume control in voice systems, real-time alerts to prevent unintended recording, or the option for anonymous profiles) could clarify the optimal conditions under which users feel confident enough to disclose personal information. This, in turn, would help reduce context-collapse concerns, foster greater trust, and further validate the effectiveness of AI-based psychotherapy platforms.

REFERENCES

[1] Ping Y. Experience in psychological counseling supported by artificial intelligence technology. Technol Health Care. 2024;32(6):3871-3888.

[2] Beg MJ, Verma M, M. VCKM, Verma MK. Artificial Intelligence for Psychotherapy: A Review of the Current State and Future Directions. Indian Journal of Psychological Medicine. 2024.

[3] Yeo YH, Clark A, Mehra M, et al. AI-Enabled Conversational Agent in Virtual Reality for Patients with Alcohol-Associated Cirrhosis. J Med Extended Reality. 2024;1(1):257-270.

[4] Miner AS, Milstein A, Hancock JT. Talking to Machines About Personal Mental Health Problems. Proc ACM Hum-Comput Interact. 2017;1(CSCW):51:1-51:19.

[5] Mansour A, Amir O, Levontin L. The Potentially Negative Effects of AI Writing Assistants on Self-Disclosure. In: The Third Workshop on Intelligent and Interactive Writing Assistants (In2Writing '24), May 11, 2024, Honolulu, HI, USA. ACM; 2024:1-4.

[6] Alanzi TM, Alharthi A, Alrumman S, et al. ChatGPT as a psychotherapist for anxiety disorders: An empirical study. Nutr Health. 2024.

[7] Kling M, Haeussl A, Dalkner N, et al. Social Robots in Adult Psychiatry. Front Psychiatry. 16:1506776.

[8] Köhler S, Guhn A, Betzler F, et al. Therapeutic self-disclosure within DBT, schema therapy, and CBASP. Front Psychol. 2017;8:2073.

[9] Marais G, McBeath A. Therapists' lived experience of self-disclosure. Eur J Qual Res Psychother. 2021;11:72-86.

[10] Lee Y, Choi S, Kim S, et al. Understanding the Impact of Long-Term Memory on Self-Disclosure in AI-Driven Health Monitoring. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. ACM; 2024:1-14.

[11] Davis JL, Jurgenson N. Context collapse: Theorizing context collusions and collisions. Inf Commun Soc. 2014;17(4):476-485.

[12] Kumar PC, Zimmer M, Vitak J. A Roadmap for Applying Contextual Integrity. Proc ACM Hum-Comput Interact. 2023;7(CSCW1):1-29.

[13] Masaviru M. Self-Disclosure: Theories and Model Review. J Culture Society Dev. 2016;18:43-47.

[14] Wester J, Siebert LC, Degen H, et al. Perceived Moral Agency of Mental Health Chatbots. Proc ACM Hum-Comput Interact. 2024;8(CSCW1):Article 133.

[15] Petronio S. Brief status report on communication privacy management theory. J Fam Commun. 2013;13(1):6-14.