Designing Human-centric AI Mental Health Chatbots : A Case Study of Two Apps

Kate Sangwon Lee
National University of Singapore

Designing Human-Centric AI Mental Health Chatbots: A Case Study of Two Apps

Kate Sangwon Lee, Janice Yeung, Arianni Kurniawati, and Dillon Tingyang Chou

Abstract Artificial intelligence (AI) mental health chatbots have become popular alternative tools in mental care by enhancing the accessibility and scalability of psychotherapy. However, integrating human-centric AI technology and mental health services poses challenges, as safety, reliability, and trust issues have not been fully addressed. This study conducted a case study with two AI mental health apps, Wysa and Youper, to explore the implications of using AI mental health chatbots as therapeutic tools to provide accessible, immediate, and personalized mental health support. The study emphasizes the importance of a human-centric AI approach by examining whether these apps are ethically aligned and reliable enough to address diverse symptoms effectively. Through a comprehensive analysis, this paper outlines design suggestions for developing human-centric AI mental health chatbots while also examining potential challenges and limitations, such as maintaining partnerships with human therapists, ensuring privacy, and the necessity for continuous improvement through feedback loops. With careful design and ethical consideration, AI mental health chatbots can significantly contribute to the mental health field by offering solutions to reduce the burdens of manual tasks of human therapists and providing accessibility to patients.

Keywords Human-centric AI · Mental health · Chatbot · Expert review

K. S. Lee (✉)
National University of Singapore, Engineering Drive 2, Singapore, Singapore
e-mail: katelee@nus.edu.sg

J. Yeung
Expedia Group, 407 St John St, London EC1V 4EX, UK

A. Kurniawati
UX UI Designer and Front-End Developer, Hong Kong, China

D. T. Chou
National Taipei University of Education, Da-an District, Taipei City, Taiwan

A. Iglesias et al. (eds.), *Information Systems for Intelligent Systems*, Lecture Notes in Networks and Systems 1255, https://doi.org/10.1007/978-981-96-1747-0_36

1 Introduction

A current study conducted in 2022 by the World Health Organization (WHO) showed that one in every eight people in the world reported experiencing a mental disorder [53], and the number has been rapidly increasing during and after the COVID-19 pandemic [19, 53, 54]. Due to the limited mobility and infection rates, COVID-19 affected a 27.6% increase in major depressive disorders and a 25.6% increase in anxiety disorders worldwide in 2020, the first year of the pandemic [54]. However, there is insufficient professional care to address these issues [19, 54], especially the difficulties raised in marginalized communities such as rural and low-income areas [8]. Furthermore, the pandemic made it more challenging to access in-person care due to limited mobility [8, 54]. The lack of professionals and scarcity of accessibility contribute to this issue, and financial burdens, discrimination, and social stigma are still some of the reasons that patients are hesitant to seek professional help [19, 37, 39, 52].

To address these challenges, virtual agents powered by conversational AI can be an alternative solution in mental health care since they enhance the accessibility, affordability, and scalability of the therapy [8, 19, 44]. AI chatbots, due to their anonymity and accessibility, can provide immediate support and have been found to mitigate mental issues effectively [8, 19, 44]. After the introduction of ChatGPT in November 2022, the utilization of natural language procession (NLP) has accelerated this trend of AI mental health chatbots in the market [12, 19, 23]. The AI mental health applications could act as patient-facing chatbots [1, 8] or back-end assistants [32], providing therapists with aid in tasks such as initial mental health assessments or insights garnered from its large language model (LLM) processing capabilities [12, 44].

However, few clear guidelines strictly govern the use of conversational AI for mental health purposes [1, 40, 44, 46]. Possible abuses or manipulations of these mental health-related technologies might lead to serious consequences [19, 44]. Although there has been a long history of mental health-related apps in the market [19, 32], there were not many that were proven clinically validated [1, 8, 19, 46]. Furthermore, not every current mental health AI chatbot in the market has been critically or empirically examined regarding safety and efficacy by checking whether the chatbots adopted human-centric approaches in using AI in the mental health domain [1, 8, 46]. To address this research gap, this study conducted a case study with two AI mental health chatbot apps to examine whether these real-world applications provided human-centric approaches. By conducting this case study, we propose how AI mental health chatbots should be designed to provide human-centric AI functionalities and enhance therapy efficacy and safety.

2 Related Work

This section begins with a brief overview of the current development of AI in mental health. It outlines (1) AI mental health apps-related research and (2) human-centric AI frameworks and guidelines.

2.1 Mental Health AI

There have been increasing numbers of commercially available AI digital applications for mental health care [19, 47]. AI has been used in various areas of mental health, such as psychiatric diagnosis [8, 12, 32, 44], self-monitoring [23], relevant content delivery about therapeutic techniques [8], and psychotherapy [5, 8, 13, 26, 39, 56]. The characteristics of Gen AI are relevant in conversational psychotherapy since it enables personalized conversations tailored to the patient's needs and symptoms [19, 24, 44, 46]. The potential of these apps lies in providing first-hand care for inclusive users since it frees worries of stigmatization, financial burdens, and physical distances [39, 47].

However, despite the prevalence of its usage, the evidence around the safety and efficacy of these tools for mental health is not fully proven [8, 19, 47]. Most of the previous studies focused on the apps' usability, efficacy, and acceptability [8]. A systematic review conducted in 2020 found that there was no statistically significant in using AI mental health apps regarding psychological well-being and anxiety [1]. A systematic review conducted in 2019 found that mobile applications for depression, anxiety, self-harm, and sleep issues only demonstrated significant positive effects for depression, with some effects observed for smoking cessation and sleep problems [51].

Also, criticism has been that these apps could generalize users' diverse situations and statuses [8, 48, 51]. Especially since there are many needs in historically marginalized user groups such as low-income or minorities, these AI mental health apps should incorporate non-biased and culturally appropriate appreciations [19]. Risks have been noticed regarding issues such as fairness and bias, accountability, privacy, and transparency [1, 19, 44]. These concerns should be addressed to meet a larger unmet need for mental health care.

2.2 Human-Centric AI Frameworks and Guidelines

Human-centric AI is a notion and scheme for building AI technologies that enhance human performance and important values such as reliability, safety, and trustworthiness [42, 43]. By taking this approach, human-centric AI technologies can help enhance human values such as rights, justice, and dignity [43]. They can be developed with values such as fairness, reliability, safety, privacy, inclusiveness, transparency, and accountability [30, 33, 41]. Various frameworks have been proposed to achieve this notion [2, 29, 43], proposing critical values such as transparency [7, 16, 27, 38], privacy [25, 45], and autonomy [15, 43]. The concept of responsible AI (RAI) [6, 28, 50] and explainable AI (XAI) [18, 29, 34, 49] have been proposed related to human-centric AI. The basic notion of explainable AI is that AI apps should ensure human control while increasing automation and provide enough explanation, transparency, and interpretability about the rationales of AI's automation [24, 41, 43]. Explanation occupies critical portions in human-centric AI since the methods around explainable AI (XAI) can help achieve human-centric AI by providing methods for users to intervene and understand AI's automated interventions [9, 14, 24, 29, 35, 49].

There are several industry human-centric AI guidelines that practitioners can use to deliver these values in human-centric AI frameworks [41, 55, 57]. The tech giants Apple, Google, and Microsoft provide human-AI interaction guidelines that emphasize interface, deployment, initial consideration, and model [2, 55, 57]. Among these industries, human-AI interaction and human-centric AI guidelines, this study particularly referred to Google's People+AI Guidebook, as shown in Table 1 in Appendix 1 [57]. The guidebook provided 23 guidelines for designing human-centric AI with detailed principles and examples. Among various guidelines, Google's guidelines were chosen since they provided comprehensive, practical, and detailed guidance with examples that practitioners can easily adopt when examining AI apps [55].

3 Method

This study conducted a case study of two generative AI mental health applications that provided chatbot functions: Wysa and Youper. We chose a case study as a method of this research since it effectively examines "a single unit for the purpose of understanding a larger class of units" [17] as we aimed to understand the current AI mental health technology's status in terms of human-centric AI frameworks and guidelines. The two apps were chosen based on researchers' search criteria using keywords such as "AI," "mental health," and "chatbot" in mobile application app stores, such as iOS Appstore and Google Play, as of March 2024. From the first search, researchers found several apps that included all these keywords in the app titles and descriptions on the app introduction pages in the app stores. Researchers downloaded those apps and examined whether the apps utilized AI features actively in the apps. From this process, the final two apps were decided as the main objectives of this study, as their main features are AI chatbots for users' mental health.

Wysa and Youper were built based on cognitive behavioral therapy (CBT), the most common and well-studied therapy method [8], so the conversations were curated based on CBT programs and on-demand support [46]. CBT refers to a set of interventions based on the premise that cognitive factors cause mental and psychological disorders [8, 10, 22, 46]. Based on CBT, the two apps adopted similar approaches that ask questions to check users' mental status and allow users to be actively involved

Table 1 Google's People+AI guidebook's 23 patterns

No	Pattern	Description
1	Determine if AI adds value	AI is better at some things than others. Make sure that it's the right technology for the user problem you're solving
2	Set the right expectations	Be transparent with your users about what your AI-powered product can and cannot do
3	Explain the benefit, not the technology	Help users understand your product's capabilities rather than what's under the hood
4	Be accountable for errors	Understand the types of errors users might encounter and have a plan for resolving
5	Invest early in good data practices	The better your data planning and collection processes, the higher quality your end output
6	Make precision and recall tradeoffs carefully	Determine whether to prioritize more results or higher quality results based on your product's goals
7	Be transparent about privacy and data settings	From initial onboarding through ongoing use, continue to communicate about settings and permissions
8	Make it safe to explore	Let users test drive the system with easily reversible actions
9	Anchor on familiarity	As you onboard users to a new AI-driven product or feature, guide them with familiar touchpoints
10	Add context from human sources	Help users appraise your recommendations with input from third-party sources
11	Determine how to show model confidence, if at all	If you decide to show model confidence, make sure it's done in a way that's helpful to your users
12	Explain for understanding, not completeness	Focus on giving your users the information they need in the moment, rather than a full run-down of your system
13	Go beyond in-the-moment explanations	Help users better understand your product with deeper explanations outside immediate product flows
14	Automate more when risk is low	Consider user trust and the stakes of the situation when determining how much to automate
15	Let users give feedback	Give users the opportunity for real-time teaching, feedback, and error correction
16	Let users supervise automation	Maintaining control over automation helps users build comfort and correct when things go wrong
17	Automate in phases	Progressively increase automation under user guidance
18	Give control back to the user when automation fails	Give your users a way to move forward even when the system fails or offers poor quality output
19	Design for your data labelers	Make sure that data labelers have well designed tools and workflows
20	Actively maintain your dataset	Maintain the quality of your product experience by proactively maintaining the quality of your data
21	Learn from label disagreements	Understand differences in how labelers interpret, and apply labels to prevent problems later on

(continued)

Table 1 (continued)

No	Pattern	Description
22	Embrace "noisy" data	The real world is messy! Expect the same from the data that you gather
23	Get input from domain experts as you build your dataset	Building partnerships with domain experts early can help reduce iterations on your dataset later on

Table 2 Profile of experts

Expert number	Gender	Occupation	Years of relevant experience (years)
E1	F	UX educator and researcher	15
E2	F	UX designer	6
E3	M	Master student in Psychology	6
E4	F	UX designer	5

in the conversation, which can reduce their anxiety around malicious cognitions and behavioral patterns. Also, the apps use natural language processing/understanding (NLP/NLU) algorithms to understand users' input and generate guidance based on users' input. Their website introduces this information about their basic algorithm techniques using NLP as they run based on conversational nodes within a decision-tree structure.

The expert group downloaded the apps and used them for two weeks to examine the features and detailed user interfaces (UIs) by applying Google's People+AI guidelines to evaluate whether they were applied to the app's features and UIs. Before and after this examination, two discussion sessions were conducted to reach a consensus among the experts. The case study was conducted with three experts in UX and one in psychotherapy. The detailed profiles of the four experts are shown in Table 2 in Appendix 2. Three experts were from HCI and UX backgrounds and had experiences in the relevant industry. One expert (E3) had both HCI and psychology expertise. Thus, his comments were relevant to examine the app's efficacy in psychotherapy and mental health exercises.

4 Findings

Wysa provides an AI chatbot, self-care resources, and exercises, such as improving self-esteem and managing anxiety. It actively involves human resources and a helpline when users need urgent help (e.g., "talk to a counselor, SOS"). The app covers various topics in mental health, including anxiety, sleep disorders, stress, and post-traumatic stress disorder (PTSD).

Youper provides an AI and CBT-based chatbot and features such as daily check-in and monitoring the progress related to various mental disorders such as anxiety, depression, and PTSD. It also gives insights, reports, and journaling interventions to help users track their progress and various tests to allow them to monitor their mental status.

4.1 Anchoring on Familiarity

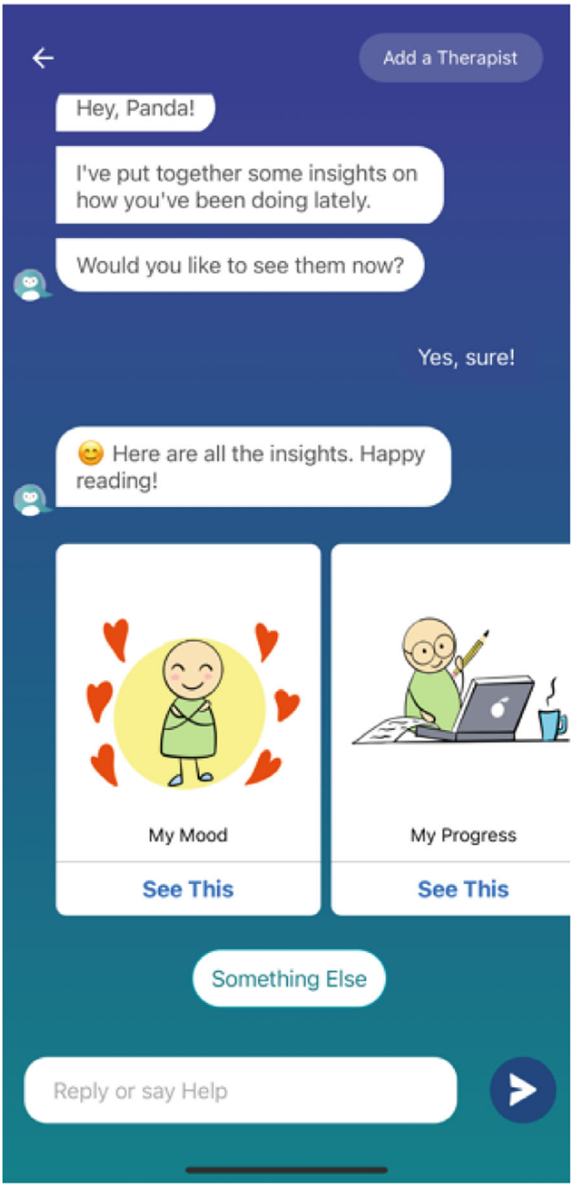
Experts agreed that Wysa provided personalized, succinct, and to-the-point answers to users without overexplaining the entire psychological background and context. Regarding "anchoring on familiarity," in the guidelines, Wysa mimics human interaction, familiar characters, and human-like wording to lower users' barriers to new technology (see Fig. 1). This approach is consistent with using common and user-friendly words and characters not only in therapeutic conversation but also in technical setting processes, such as setting notification time (e.g., asking time in general wordings: "I will reach out to you at 10:00 pm"). Compared to Youper, Wysa strategically used more familiar approaches such as character.

4.2 Utilizing Human Resources

Regarding the guideline "adding context from human resources," Wysa actively incorporated human resources into various pages. It provided relevant links to human experts on the main page and chatting pages to leverage human resources and show accessibility to them whenever the user found it necessary. Youper also distinguished its role from that of a human therapist and was transparent about being an AI-driven tool, not a substitute for professional mental health care. This helped to set user expectations appropriately. The app explicitly states in its introduction, "Youper is designed to provide support and is not a replacement for professional therapy."

However, it was found that the app did not introduce proper human resources depending on users' various symptoms and scenarios. This might lead to a lack of human oversight, especially in serious situations where professional intervention is necessary. In the current apps, a user in severe distress might not have an immediate way to connect with a human therapist, relying solely on the AI's support, which may not be sufficient. Furthermore, users' symptoms covered in the apps varied, such as anxiety, sleep disorder, and PTSD; however, the links to human support stayed the same and generic in every conversation.

Fig. 1 Wysa recommends relevant content based on users' sharing



Hey, Panda!

I've put together some insights on how you've been doing lately.

Would you like to see them now?

Yes, sure!

😊 Here are all the insights. Happy reading!

4.3 *Personalization*

Generally, experts found that Youper performed better during the chatbot's interaction than Wysa due to the simple UI focused on the chatbot and its intelligence during the conversations. For example, the conversations of Youper's chatbot were built on user responses and adapted answers dynamically and in more personalized ways. However, as mental health issues vary across different individuals, a targeted and personalized approach was lacking for both apps. Regarding the "determining if AI adds value" guideline, the experts found that the two apps did not provide enough value from the AI technology, such as providing personally customized or tailored recommendations based on different individual issues. Most of the responses from Wysa were found to be generic and not detailed, and they were not tailored to users' situations or various emotional statuses. For example, when a user input, "the therapy is not helping," Wysa only gave generic responses such as, "I'm sorry about that," "Do help me understand. Tell me what wasn't working for you," and provided general basic prompts as possible users' responses to the app's chat such as "I need more help, I feel unheard, and I want something else." The process of conversation was not specific nor personalized regarding the user's various symptoms, which did not fully leverage AI technology's advantages.

Furthermore, both chatbots did not use users' previous history in the conversation, which made the chat more generic and time-consuming. This was the point of a significant difference between human and AI therapists since human therapists mostly start conversations based on the previous therapy history and use the context actively. This helped effective therapy, which was lost in the interaction between AI therapists.

4.4 Explainability

Regarding the guideline, "let users give feedback," both Wysa and Youper occasionally asked for users' feedback during the conversations to ensure their performance level worked well, such as "May I know how I'm doing so far?" (Wysa). From the questions, the chatbot received real-time feedback from users about their therapy's effectiveness. However, in terms of the guideline, "go beyond in-the-moment explanations," the apps did not share any transparent explanations about how it would reflect users' feedback to improve further therapy.

There was also a lack of explainability of the therapy method; even though the apps introduced a CBT approach during the onboarding phase, there was no comprehensive or personalized coaching on using the chatbots based on the approach during usage. The user had to try tapping on different parts of the app to check whether there were more useful features for them. The app collected information about each user during the onboarding phase, so there would possibly be personalized and targeted features based on the user's input.

For Youper, explainability was also the app's weak point; it only emphasized the technologies but did not fully explain how they used users' collected data. There was little coaching about the methods or approaches when the user landed on the app's main interface. Regarding the guidelines, "set the right expectations" and "explain the benefit, not the technology," the app could provide more clear and detailed information about how it can help and benefit users with which methods during usage.

4.5 Comprehensive Onboarding for Deeper Explanations

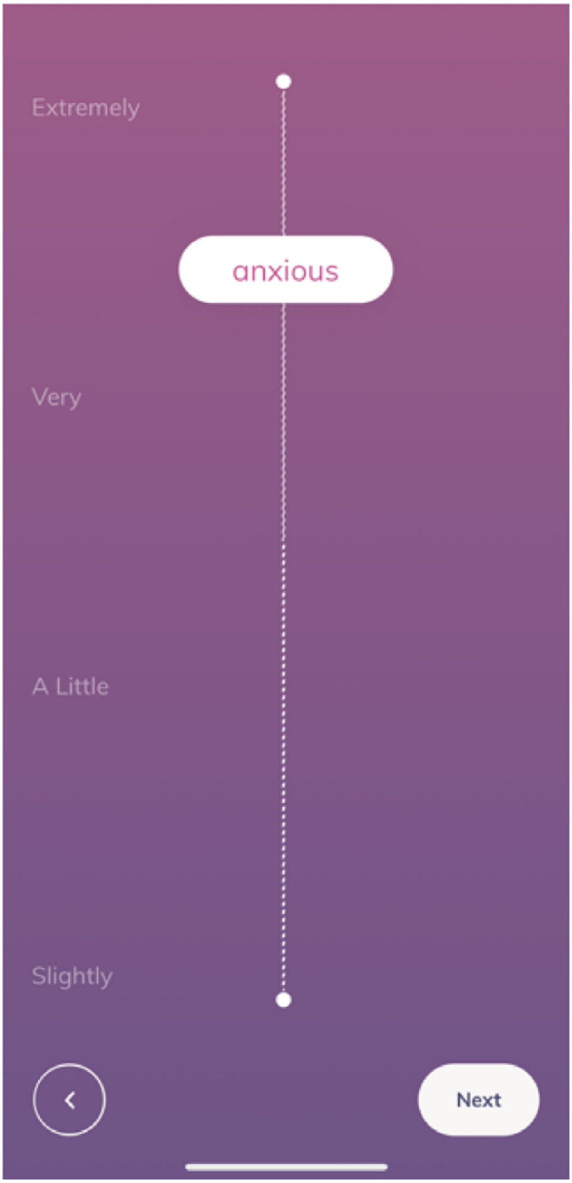
In terms of the guideline, “go beyond in-the-moment explanations,” The experts found that Youper initially provided a longer and more detailed onboarding process than Wysa, which allowed users to share their current status about their mental health and interests and receive personalized content based on the inputs (see Fig. 2). During the onboarding phase, the app showed how it incorporates CBT and how CBT would work effectively in improving users’ mental health by changing negative thought patterns into positive emotions and behaviors. Also, the app asked about users’ familiarity level with CBT, and based on the response, the app differentiated the content tailored to users’ input.

4.6 Lack of Transparency in Privacy Setting

Regarding the guideline, “be transparent about privacy and data settings,” both apps emphasized privacy’s importance, but they did not provide enough explanations or interventions to explain and allow users to control private data settings. During onboarding, Wysa stated, “Our conversations are private and anonymous, so there is no login. Just choose a nickname, and we’re good to go.” However, it did not give more detailed explanations or accounts of how it uses the information it could collect from users.

Similarly, during the initial onboarding phase, Youper also emphasized the importance of privacy (e.g., “privacy is our priority”). However, the app pages did not provide any relevant links or documentation about how users manage data settings or privacy settings and how the app uses collected data from users. Also, it did not provide an option to opt out of collecting cookies or data settings. In Youper’s settings menu, the app allowed users to download the chatting history and provided a feature to add one more layer of data security by enabling it, but it did not provide a detailed explanation about how the app enhanced security by opting into the option.

Fig. 2 Youper's onboarding step allows users to input their current feelings



5 Discussion

5.1 *Leveraging Human Resources*

The expert suggested that clarifying the limitations and recommending the benefits of using these technologies in the services should be necessary in AI mental health apps. More importantly, it should be clearly noted that the current level of AI mental health chatbots cannot replace human therapists in terms of the depth and professionalism of the conversation [39, 46]. Instead, experts found that both apps can work as supplementary tools for psychotherapy, such as mood-tracking or journaling purposes. Still, they cannot fully function or replace human therapists when users need a deeper or more advanced level of psychotherapy [39]. According to the model of helping skills in psychotherapy, there are three stages of psychotherapy in real practice: (1) exploration for self-awareness, (2) finding patterns or insights, and (3) actions for changes [20, 21]. Those apps can be helpful for the first stage, which is helping users explore their current status and emotional trends to enhance self-awareness. The human therapist can help patients move toward the second and third levels by having deep and advanced conversations. The current AI chatbots cannot achieve this level of personalized and sophisticated conversation or support [8, 19, 46].

Miner et al. [31] introduced four approaches to care provision with AI: (1) human only, (2) human delivered, AI informed, (3) AI delivered, human supervised, and (4) AI only. All these different approaches can tackle different implications and objectives regarding the advancement of technologies. Regarding the current technology level, a second approach can be adopted in the AI mental health chatbots. Also, regarding the fundamental limitation of AI as not being a rational and moral agent with an autonomous agency [39], the fourth approach (AI only) is not yet a desirable concept for advanced therapy due to ethical issues [39]. In therapy, it is critical to maintain the therapeutic relationship between the human therapist and patients during the therapy history [31, 39], and this relationship is built based on the rationale that the agent is sentient, genuine, and empathetic. However, AI chatbots cannot fully function by mimicking human intervention due to their limited authentication and lack of agency in ethical considerations, which might cause serious problems, such as the illusion of users—false beliefs and wrong expectations [39]. The main interventions of current AI mental health chatbots are checking users' moods and showing progress in users' mental health. This check-in occupies only a very small part of human therapy [21]. However, this technology certainly provides merits in tracking and self-monitoring users' daily status and guiding users about useful methods to maintain mental health. Active and relevant collaboration between human and AI therapists can lower barriers to psychotherapy and provide scalable support [8, 40, 46].

5.2 Explainable AI

The lack of explainability of the two apps was also criticized since the chatbots did not provide enough explanations as to why each step of the conversation is important to enhance users' mental health and how every step was useful in the care. Also, Wysa frequently tapped into users' feedback on its performance, though it did not show how it would reflect users' feedback for further adjustment. For example, when a user replied that the conversation was not effective, there were no clear answers about how it would be improved to satisfy users.

Furthermore, the apps adopted CBT and clarified that their approaches are based on this method during the onboarding. However, the apps did not clearly show the overall trajectory of the therapy and how the users' status is improved based on the method. Thus, users might feel passive and do not have enough autonomy about how the apps' interventions effectively enhance their mental health conditions. Explainability enhances users' trust and autonomy when using AI apps [4, 14, 24, 29, 35]. AI mental health apps can achieve trust and reliability by providing enough rationales for each therapy step and users' feedback [55].

5.3 Privacy in Self-disclosure

In terms of privacy management, even though the two apps emphasized the importance of maintaining users' privacy and private data on the app pages, they did not explicitly show the methods and detailed interventions to manage it inside the applications; for example, Youper did not give the option to opt out of specific cookies and data settings. Privacy is critical in psychotherapy since certain conversations from patients about sensitive topics can be specific, ethically problematic, or serious to be shared [32], such as an attempt to trauma, sexual history, and thoughts of self-harm [31]. If this conversation can be used to train data or marketing purposes, this might bring ethical issues and, more seriously, intrusive consequences to whom is at mentally high risk. In Youper, there is an option for "scientific research," which allows the app to collect data for scientific mental health research, and it declares that "no identifiable personal information, conversations or messages will ever be used in the research." However, there was no clear explanation of what this identifiable personal information indicates and how the app can maintain the confidentiality of users' therapy.

In human therapy, the protection of privacy is ethically and legally established [36], such as the Health Insurance Portability and Accountability Act (HIPAA) regulations of 1996 [19, 32]. The adaptation of existing policy can improve privacy management in AI mental health apps. As in in-person therapy, in the apps, users should have autonomy and be given explanations about private data control and management [8, 45, 55]. The standards and legitimate process for privacy policies in AI mental health apps should be established to protect users' privacy and authority in data management [19, 32].

5.4 Design Suggestions

The experts discussed possible design suggestions from the analysis, such as below:

- **Initial interaction**: Provide comprehensive onboarding to guide users on how to use this new exciting technology and give examples of how the chatbot experience can deliver better or new value compared to a real-life therapist.
- **Initial interaction**: Be transparent about what the chatbot can or cannot do and actively leverage human resources—refer to real-life human therapists when conversations go beyond the chatbot's capabilities.
- **Initial interaction and over time**: Provide options to opt out of collecting private data (e.g., cookies) and user-friendly explanations about data and private information management.
- **Over time**: Empower the partnership between human and AI mental health chatbots and provide proper and timely human interventions when users' current symptoms and levels need human resources.
- **Over time**: Provide personalized approaches based on users' different symptoms and users' previous conversation data.

Furthermore, advanced ideas are discussed: as in real therapy sessions with human therapists, there would be richer information conveyed from patients' facial expressions, tones, manners of voice, and body language [3, 11, 32]. The current AI chatbots only provide typing as a method of input, which does not convey full context. For future improvement, we propose that it can incorporate voice and video inputs so AI can detect richer information about users' emotional status by using voice, faces, and body expressions.

5.5 Limitation and Future Work

The potential for future work includes conducting user studies for both sides—patients and therapists, as the current study was conducted only with experts. A suggested future study could involve a longitudinal user study to test the effectiveness of AI mental health chatbots with real users over a certain period of time. It is recommended in this paper that a partnership between human and AI therapists should be explored to enhance the efficacy and safety of therapy, although the details of this partnership were not fully described. Future work could involve conducting studies to explore and test different types of partnerships with real therapists and patients and develop effective workflow between human and AI therapists. Possible future work can also incorporate various types of AI mental health chatbots to increase the comprehensive applicability of the findings.

6 Conclusion

This study conducted a case study with two real-world AI mental health chatbots with human-centric AI guidelines to examine the apps' current level of therapy efficacy and holistic user experience regarding human-centric AI frameworks and related values. The results show that the current levels of these chatbots complement some points of guidelines, such as anchoring on familiarity, allowing users to give feedback, and adding context from human resources. However, some points, such as leveraging human resources properly, personalization, transparency about privacy settings, and showing enough rationales for their responses for better explainability, were not achieved enough in the apps. This study recommends that for the future improvement of AI mental health chatbots, interventions should be incorporated to harness human therapy effectively and protect users' privacy settings, and clear explanations should be shared for users' understanding and respect for autonomy. The contribution of this study can be summarized as (1) examining current AI mental health chatbots with human-centric guidelines and analyzing how their designs were aligned with the guidelines and (2) proposing human-centric AI mental health chatbot design suggestions.

Appendix 1

See Table 1.

Appendix 2

See Table 2.

References

1. Abd-Alrazaq AA, Rababeh A, Alajlani M, Bewick BM, Househ M (2020) Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. J Med Internet Res 22(7):e16021
2. Amershi S, Weld D, Vorvoreanu M, Fourney A, Nushi B, Collisson P, Suh J, Iqbal S, Bennett PN, Inkpen K, Teevan J, Kikin-Gil R, Horvitz E (2019) Guidelines for human-AI interaction.

In: Proceedings of the proceedings of the 2019 CHI conference on human factors in computing systems. association for computing machinery

3. Amichai-Hamburger Y, Klomek AB, Friedman D, Zuckerman O, Shani-Sherman T (2014) The future of online therapy. Comput Hum Behav 41:288–294
4. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Garcia S, Gil-Lopez S, Molina D, Benjamins R, Chatila R, Herrera F (2020) Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 58:82–115
5. Bendig E, Erb B, Schulze-Thuesing L, Baumeister H (2022) The next generation: chatbots in clinical psychology and psychotherapy to foster mental health–a scoping review. Verhaltenstherapie 32(Suppl. 1):64–76
6. Benjamins R, Barbado A, Sierra D (2019) Responsible AI by design in practice. arXiv preprint arXiv:1909.12838
7. Betzing JH, Tietz M, vom Brocke J, Becker J (2020) The impact of transparency on mobile privacy decision making. Electron Mark 30:607–625
8. Boucher EM, Harake NR, Ward HE, Stoeckl SE, Vargas J, Minkel J, Parks AC, Zilca R (2021) Artificially intelligent chatbots in digital mental health interventions: a review. Expert Rev Med Devices 18:37–49
9. Brdnik S (2023) GUI design patterns for improving the HCI in explainable artificial intelligence. In: Proceedings of the companion proceedings of the 28th international conference on intelligent user interfaces. Association for Computing Machinery
10. Butler AC, Chapman JE, Forman EM, Beck AT (2006) The empirical status of cognitive-behavioral therapy: a review of meta-analyses. Clin Psychol Rev 26(1):17–31
11. Castellano G, Kessous L, Caridakis G (2008) Emotion recognition through multiple modalities: face, body gesture, speech. Springer Berlin Heidelberg
12. Cheng SW, Chang CW, Chang WJ, Wang HW, Liang CS, Kishimoto T, Chang JPC, Kuo JS, Su KP (2023) The now and future of ChatGPT and GPT in psychiatry. Psychiatry Clin Neurosci 77(11):592–596
13. Das A, Selek S, Warner AR, Zuo X, Hu Y, Keloth VK, Li J, Zheng WJ, Xu H (2022) Conversational bots for psychotherapy: a study of generative transformer models using domain-specific dialogues
14. Dwivedi R, Dave D, Naik H, Singhal S, Omer R, Patel P, Qian B, Wen Z, Shah T, Morgan G (2023) Explainable AI (XAI): core ideas, techniques, and solutions. ACM Comput Surv 55(9):1–33
15. Evers C, Kniewel R, Geihs K, Schmidt L (2014) The user in the loop: enabling user participation for self-adaptive applications. Futur Gener Comput Syst 34:110–123
16. Felzmann H, Fosch-Villaronga E, Lutz C, Tamò-Larrieux A (2020) Towards transparency by design for artificial intelligence. Sci Eng Ethics 26(6):3333–3361
17. Gerring J (2004) What is a case study and what is it good for? Am Polit Sci Rev 98(2):341–354
18. Ghai B, Liao QV, Zhang Y, Bellamy R, Mueller K (2021) Explainable active learning (xal) toward ai explanations as interfaces for machine teachers. In: Proceedings of the ACM on human-computer interaction, 4, CSCW3, pp 1–28
19. Hamdoun S, Monteleone R, Bookman T, Michael K (2023) AI-based and digital mental health apps: balancing need and risk. IEEE Technol Soc Mag 42(1):25–36
20. Hill CE (2020) Helping skills: facilitating exploration, insight, and action. American Psychological Association
21. Hill CE, Stahl J, Roffman M (2007) Training novice psychotherapists: helping skills and beyond. Psychoth Theo Res Pract Train 44(4):364
22. Hofmann SG, Asnaani A, Vonk IJ, Sawyer AT, Fang A (2012) The efficacy of cognitive behavioral therapy: a review of meta-analyses. Cogn Ther Res 36:427–440
23. Javaid M, Haleem A, Singh RP (2023) ChatGPT for healthcare services: an emerging stage for an innovative perspective. BenchCouncil Trans Benchmarks, Stand Eval 3(1):100105
24. Joyce DW, Kormilitzin A, Smith KA, Cipriani A (2023) Explainable artificial intelligence for mental health through transparency and interpretability for understandability. NPJ Digit Med 6(1):6–6

25. Kim TW, Routledge BR (2018) Informational privacy, a right to explanation, and interpretable AI
26. Lee J, Lee D, Lee J-G (2024) Influence of rapport and social presence with an AI psychotherapy chatbot on users' self-disclosure. Int J Human–Comput Interact 40(7):1620–1631
27. Liao QV, Subramonyam H, Wang J, Wortman Vaughan J (2023) Designerly understanding: information needs for model transparency to support design ideation for AI-powered user experience
28. Liao QV, Sundar SS (2022) Designing for responsible trust in AI systems: a communication perspective
29. Liao QV, Varshney KR (2021) Human-centered explainable AI (xai): from algorithms to user experiences. arXiv preprint arXiv:2110.10790
30. Lu Q, Zhu L, Xu X, Whittle J, Zowghi D, Jacquet A (2024) Responsible AI pattern catalogue: a collection of best practices for AI governance and engineering. ACM Comput Surv 56(7):1–35
31. Miner AS, Shah N, Bullock KD, Arnow BA, Bailenson J, Hancock J (2019) Key considerations for incorporating conversational AI in psychotherapy. Front Psych 10:746
32. Olawade DB, Wada OZ, Odetayo A, David-Olawade AC, Asaolu F, Eberhardt J (2024) Enhancing mental health with artificial intelligence: current trends and future prospects. J Med Surg Public Health 100099
33. Ozmen Garibay O, Winslow B, Andolina S, Antona M, Bodenschatz A, Coursaris C, Falco G, Fiore SM, Garibay I, Grieman K, Havens JC, Jirotka M, Kacorri H, Karwowski W, Kider J, Konstan J, Koon S, Lopez-Gonzalez M, Maifeld-Carucci I, McGregor S, Salvendy G, Shneiderman B, Stephanidis C, Strobel C, Ten Holter C, Xu W (2023) Six human-centered artificial intelligence grand challenges. Int J Human–Comput Interact 39(3):391–437
34. Qian K, Popa L, Sen P (2019) Systemer: a human-in-the-loop system for explainable entity resolution. Proc VLDB Endow 12(12):1794–1797
35. Rjoob K, Bond R, Finlay D, McGilligan V, Leslie SJ, Rababah A, Iftikhar A, Guldenring D, Knoery C, McShane A, Peace A (2021) Towards explainable artificial intelligence and explanation user interfaces to open the 'black box' of automated ECG interpretation. Springer International Publishing
36. Roberts LW (2016) A clinical guide to psychiatric ethics. American Psychiatric Pub
37. Samari E, Teh WL, Roystonn K, Devi F, Cetty L, Shahwan S, Subramaniam M (2022) Perceived mental illness stigma among family and friends of young people with depression and its role in help-seeking: a qualitative inquiry. BMC Psych 22(1):107
38. Schmidt P, Biessmann F, Teubner T (2020) Transparency and trust in artificial intelligence systems. J Decis Syst 29(4):260–278
39. Sedlakova J, Trachsel M (2023) Conversational artificial intelligence in psychotherapy: a new therapeutic tool or agent? Am J Bioeth 23(5):4–13
40. Sharma A, Lin IW, Miner AS, Atkins DC, Althoff T (2023) Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nat Mach Intell 5(1):46–57
41. Shneiderman B (2020) Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Trans Interact Intell Syst 10(4), Article 26
42. Shneiderman B (2022) Human-centered AI. Oxford University Press
43. Shneiderman B (2022) Human-centered AI: ensuring human control while increasing automation. In: Proceedings of the proceedings of the 5th workshop on human factors in hypertext. Association for Computing Machinery
44. Singh OP (2023) Artificial intelligence in the era of ChatGPT-opportunities and challenges in mental health care. Indian J Psych 65(3):297–298
45. Stahl BC, Wright D (2018) Ethics and privacy in AI and big data: implementing responsible research and innovation. IEEE Secur Priv 16(3):26–33
46. Thieme A, Hanratty M, Lyons M, Palacios J, Marques RF, Morrison C, Doherty G (2023) Designing human-centered AI for mental health: developing clinically relevant applications for online CBT treatment. ACM Trans Comput-Human Interact 30(2):1–50

47. Timmons AC, Duong JB, Simo Fiallo N, Lee T, Vo HPQ, Ahle MW, Comer JS, Brewer LC, Frazier SL, Chaspari T (2023) A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. Perspect Psychol Sci 18(5):1062–1096
48. Van Ameringen M, Turna J, Khalesi Z, Pullia K, Patterson B (2017) There is an app for that! the current state of mobile applications (apps) for DSM-5 obsessive-compulsive disorder, posttraumatic stress disorder, anxiety and mood disorders. Depress Anxiety 34(6):526–539
49. Wang D, Yang Q, Abdul A, Lim BY, Assoc Comp M (2019) Designing theory-driven user-centric explainable AI. Assoc Comput Mach
50. Wang Q, Madaio M, Kane S, Kapania S, Terry M, Wilcox L (2023) Designing responsible AI: adaptations of UX practice to meet responsible AI challenges
51. Weisel KK, Fuhrmann LM, Berking M, Baumeister H, Cuijpers P, Ebert DD (2019) Standalone smartphone apps for mental health—a systematic review and meta-analysis. NPJ Digit Med 2(1):118
52. Wijeratne C, Johnco C, Draper B, Earl J (2021) Doctors' reporting of mental health stigma and barriers to help-seeking. Occup Med 71(8):366–374
53. World Health Organization (2022) Mental disorders
54. World Health Organization (2022) Mental health and COVID-19: early evidence of the pandemic's impact: scientific brief, 2 March 2022, World Health Organization
55. Wright AP, Wang ZJ, Park H, Guo G, Sperrle F, El-Assady M, Endert A, Keim D, Chau DH (2020) A comparative analysis of industry human-AI interaction guidelines. arXiv preprint arXiv:2010.11761
56. Xu B, Zhuang Z (2022) Survey on psychotherapy chatbots. Concurr Comput Pract Exp 34(7):e6170
57. Yildirim N, Pushkarna M, Goyal N, Wattenberg M, Viégas F (2023) Investigating how practitioners use human-ai guidelines: a case study on the people+ai guidebook