Review of AI-Based Mental Health Apps

Abeer Alotaibi
Lancaster University

Corina Sas
Lancaster University

The last decade has seen a significant growth of HCI research in mental health technologies while AI raises both challenges and opportunities to better support symptom identification or personalization of interventions. There has been also a growth of commercial AI-based mobile apps for mental health. Despite emerging HCI work on reviewing mental health apps, those that are AI-based have received limited attention. To address this gap, we report a functionality review of 13 such apps selected from 127 apps from the Apple Store. Findings indicate that apps support functions such as tracking and detecting emotions and moods, providing recommendations for therapy and well-being interventions, and supporting talking therapy through conversational agents powered by NLP models. A critical finding is apps' limited support for AI literacy and explainability, as well as limited consideration for ethical concerns regarding personal data, its reliability, and algorithmic biases. Our paper concludes with three design implications for AI-based mental health apps towards developing conversational agents to support CBT interventions based on tracked multimodal data, addressing the ethics of NLP biases, and user exploration of AI-based models and their XAI.

AI, ML, Mental health, Explainable AI, Mood, Emotion, Mobile apps, Conversational agents.

1. INTRODUCTION

The last decade has seen significant growth of HCI research in mental health technologies and the potential of lowering the barriers to accessing traditional psychotherapeutic care. From mobile apps (Qu et al., 2020) and wearable devices (Alfaras et al., 2020) to social media (Saha et al., 2019) or mental health platforms (Richards et al., 2023), most of these technologies leverage self-tracking, usually of emotional or behavioural aspects (Søgaard etal., 2019) with less focus on interventions for mental health. Among the latter, the most common ones are computerized cognitive behavioural therapy (CBT) (Thieme et al., 2022) mindfulness-based interventions (Terzimehić et al., 2019), or biofeedback ones (Miri et al., 2020). Research has also started to highlight ethical challenges pertaining to vulnerable users' data privacy and the clinical validity of the technology-based intervention (Qu et al., 2020; Sanches et al., 2019;). While HCI research on mental health has developed rather independently from the work on human-AI interaction, emerging efforts have attempted to intersect them with a focus on AI-based technologies for mental health. Works on AI and mental health have focused mostly on detection and diagnosis and less so on personalised interventions, leveraging data from mobile phones, social media or health records (Thieme et al.2020). The main challenges in this area include access to quality datasets, the ethics of personal data, and ML-biases (Harrington et al., 2022Thieme et al., 2020). Additional challenges include users' limited understanding of and trust in these technologies (Liao et al., 2022). With explainable AI (XAI) for mental health has been limitedly explored on previous work (Chalabianloo et al., 2022; Zajac et al., 2023; Kim et al., 2022). Apart from academic research in this space, there has also been a growth of commercial AI-based mobile apps for mental health which have been reviewed. While such studies have called for improvement of the validity of these apps (Bowie-DaBreo et al., 2020; Qu et al., 2020; Søgaard et al., 2019) there has been limited exploration of how AI supports their features or raise additional ethical concerns. Particularly, this paper's contribution is to fill the gap in current research, which is the limited exploration of how AI can be ethically integrated into the design and user experience of mobile apps for mental health, contributing to both the AI and HCI communities. The study investigates the ethical challenges of AI-based mobile apps, emphasising interventions as a key functionality. It also addresses design implications for developing conversational agents to support CBT interventions using multimodal data, addressing NLP biases, and facilitating user exploration of AI-based models and their XAI features. The contribution lies in advancing knowledge and practices within the ML and HCI communities, improving the validity and ethical design of AI-based mental health apps. To address this gap, we reviewed the 13 most popular AI-based mental health apps on the Apple Store. Users' ratings of these apps are likely to relate to their features, some of which are likely to be grounded in novel design knowledge which we aim to identify and articulate.

We focus on the following research questions: (i) What are the key functionalities of AI-based mental health apps? (ii) What are the main ethical challenges associated with AI-based mental health apps? (iii) How do these apps support users' understand of their AI models and outputs, i.e. XAI?

2. LITERATURE REVIEW

Our literature review aims to combine various interdisciplinary sources to create inclusive framework of essential skills and competencies for Human-Computer Interaction (HCI) researchers working on the intersection of mental health technologies, AI, and AI literacy.

2.1 HCI Research on Mental Health

HCI research on mental health has grown significantly over the last decade with a major focus on automated diagnostic and self-tracking leveraging users' posts on social media platforms (Saha et al., 2019), as well visual or audio interfaces, wearables or biosensing devices (Chalabianloo et al., 2022) and less so on interventions or treatments (Sanches et al., 2019; Søgaardet al., 2019) as computerised cognitive behavioural therapy (CBT) (Grové, 2021) biofeedback (Umair et al.,2021). Sanches and colleagues' (2019) systematic review of HCI research on affective health also highlighted that. only one-third of the 139 reviewed papers reflect on ethical concerns, related mostly to the principles of autonomy and non-maleficence. From ethics lens, this review also highlights the issues regarding automatic diagnosis provided without therapeutic support, ownership of sensitive mental health personal data, and increased vulnerability of users living with mental health conditions. Important findings highlighted that despite the many available mental health technologies such as mobile apps, their evidence-based is limited hence their validity remains a concern (Khazza et al., 2018). The assessment validity of e-mental therapies generally is usually limited (Sanches et al., 2019).

In contrast to technologies developed in academia, a limited strand of HCI work has focused on commercial mobile apps for mental health.. Noticeable here is a review study of 29 top rated apps for depression, selected from 482 apps on marketplace (Qu et al., 2020). Findings indicate functionalities such as tracking moods, thoughts, behaviours, or depression symptoms, and those focused on interventions such as diaries, psychoeducation, mindfulness, positive behaviour. Authors also highlighted ethics concerns of these apps such as limited provision of privacy and safety policy to protect highly vulnerable users such as children, limited clinical input into apps'design needed to ensure their validity. The study concludes with recommendations to stronger personalization of interventions based on tracked complementary multimodal data, and for mitigation against harm. An earlier review of 353 apps for depression highlighted the need for improving their clinical validity, fidelity of the treatment and the general safety so that they better align with the NICE guidelines for depression (Bowie-DaBreo et al., 2020).

2.2 HCI Research on AI

Emerging research over the last years, has also focused on the integration of AI and ML topics in HCI and interaction design research agenda with a focus on the main challenges and ethical concerns (Loi et al., 2019). In their review of AI in the wild, Zajac and colleagues (2023) identified key sociotechnical challenges of human-AI interaction due to the not trivial to understand and design with ML capabilities, required interdisciplinary expertise and large training data sets, unpredictable outcomes, and new modes of interactions through which users ongoingly improve the algorithms. Such challenges impact both designers working in HCI and users of AI or ML based systems. An empirically rich study in Indian context where AI tends to be perceived as aspirational (Kapania et al., 2022) indicated that users' attitudes of towards AI include faith, forgiveness, self-blame and gratitude leading to increased vulnerability. Authors suggest the need to adjust AI authority through responsible AI approaches and measures of success.

The integration of AI in user-facing technologies faces challenges due to users' misconceptions and limited knowledge, leading to ineffective use as well as misguided regulation (Long and Magerko, 2020). To address these issues, AI literacy has emerged as a crucial concept, emphasising the need to equip users with competencies to understand and engage effectively with AI (Druga et al., 2022). HCI research has focused on xplainable AI (XAI) to enhance users' understanding and trust (Balkir et al., 2022) in how AI systems process data and make decisions (Mohseni et al., 2021) or judgments (Ehsan et al., 2019).

Ehsan and colleagues (2021) advanced the concept of XAI systems as being socially situated rather than algorithm centric, and proposed the social transparency perspective that accounts for socio-organizational context of such systems. Authors suggested three levels of context made visible through social transparency: technological, decision-making and organizational supporting tracking AI performance, confidence in decision-making, and accountability, respectively. They also proposed four design features: what such as AI-based actions or decisions, why such as rationale underpinning the decision, who namely the user and their organizational role, and when such as timing of the decision. Apart from such broad exploration of XAI, other strand of work has focused on comparing existing approaches or developing new ones.

One illustration of the former, is Liao and colleagues (2022) who explored the most common AI approaches used in recommender systems, whose findings indicate users' increase trust in collaborative filtering which matches users with

similar preferences, rather than matching users with product features (content-based filtering) or broader demographic characteristics (demographic filtering) there scholars developed guidelines for the development of artificial agents and their intelligibility (Cila, 2022) or novel XAI approaches aimed to support users identity bugs in AI agents, i.e., After-Action Review for AI (AAR/AI) (Roli Khanna et al., 2022). Its valuation shows that this approach support users identify more precisely more bugs, with authors concluding that labels embedded in the AAR/AI may assist users building abstract domain-knowledge from specific bug instances.

To summarise, HCI research on human-AI interaction highlighted the need for interdisciplinary expertise to address the sociotechnical challenges of AI while harnessing its affordances. Main challenges relate to the large data sets needed for such systems, and users' limited trust and understanding of their outputs. Emerging work has also explored AI techniques that best support XAI, such as those for recommender systems. The studies also emphasises the need for consumers to comprehend knowledge representations and reasoning processes of AI, and suggests incorporating explainability components in design to enhance user knowledge. The concept of social transparency in XAI systems is introduced, considering the socio-organizational context and proposing design features to support user faith.

2.3 HCI Research on AI-Based Technologies for Mental Health

While the HCI research on mental health technologies and human-AI interaction has progressed rather separately, recent efforts have begun to explore their intersection. A landmark study by Thieme et all (2020) focused on a systematic review of 54 HCI papers on machine learning (ML) for mental health, demonstrating the value of ML in detecting, diagnosing, and treating mental health conditions, including identification of symptoms and risks factors, condition prediction, and personalised interventions. The study also identified the limited use of speech and conversational agents in mental health ML applications, highlighting potential areas for exploration (Straw and Callison-Burch, 2020).

Few papers explored the potential of just in time adaptive interventions informed for instance by current emotional states or by logged data captured by mobile apps. Authors also highlight the challenges of creating or accessing sufficiently large, high quality data sets which represent the diversity of the target populations while accounting for the sensitive personal data of vulnerable users to ensure inclusive, non-harming, responsible, fair and robust AI-based technologies. Additional ethical challenges include those concerning the AI-based decision making, its accountability, risk for biases and malicious intents. While Sanches and colleagues' systematic review (2019) has not focused explicitly on AI-based systems, it highlighted the use of ML for recommending interventions and generating reminders for engaging with them, based on tracked data (Qu et al., 2020) alongside the ethical challenges of developing AI-based systems for mental health. Since these reviews, further work has focused on AI-based systems for mental health such as MindScope, a mobile app using ML algorithms such as multiple decision trees for predicting stress level from users' interaction with the phone capturing: social activity, i.e., calls, ambient noise, change of location, physical activity, i.e., walking, running and their duration, sleep activity, i.e., screen use during night hours, and phone use, i.e., screen status and unlocked duration, apps' frequency and duration of use (Simard et al , 2021). The app also employs XAI techniques to explain how the stress level was predicted.

A study by Harrington et al. (2022) explored the use of conversational voice assistants like Google Home for health information search among low-income Black older adults. The findings highlighted usability challenges and limited accessibility for these users (Cheng et al., 2019; Razak et al., 2010). Chalabianloo et al. (2022) used Shapley Additive exPlanations (SHAP) to explain ML model outcomes in a study involving wearable biosensors. Their findings showed the value of SHAP visualizations in understanding stress detection models. Zajac et al. (2023) emphasised the challenges of designing socio-technical systems for clinician-facing AI technologies. Grové (2021) and Khazaal et al. (2018) emphasised the importance of involving clinicians and users in the design of ML-based systems, distinguishing between explainable outputs with small, labelled datasets and data-driven systems with better performance but limited explainability. Their findings also highlight three types of support provided by ML- based systems in clinical settings: aiding clinicians' decision-making, facilitating clinicians' prioritization, and automating clinical tasks. They argue that HCI research should focus on interdisciplinary collaboration and the design of sociotechnical systems that support explainability using existing frameworks (Jain and Agarwal, 2017). To conclude, most of HCI research on AI and mental health has targeted less personalised interventions and mostly detection and diagnosis usually of depression based on data from mobile phones, social media, health records using mostly supervised rather then unsupervised machine learning, such as NLP for anaylyis ofonline communication (Nikiforos, Tzanavaris, & Kermanidis, 2020).

3. METHOD

To identify AI-based mental health, we conducted a search in winter 2022 on Apple Store using combined keywords with the first being one of the following “mental health”, “emotion”, “mood” and the second term being one of the following: “artificial intelligence”, “AI”, “machine learning”. “ML”, “chatbot”. Figure 1 shows the PRISMA diagram of this selection process. From the initial 127 apps, we removed 32 duplicates apps and 9 apps identified as irrelevant to mental health *such as entertainment, design, or sport apps*. We also excluded 77 apps with less than 25 raters and average ratings less than 4 out of 5. We also excluded 1 app not available in the UK store, leading to a final set of 8 commercial apps 5 of which being also available on Google Play. In addition, we also looked for AI-based mental health apps mentioned in academic papers and at the same time available on Apple Store but possibly not returned by our initial search. For this, we searched on Google Scholar using combined keywords, with the first being one of the following: “digital mental health”, “chatbot” and the second term being one of the following: “artificial intelligence”, “AI”, “machine learning”, “ML”. This search returned 5 apps also available on Apple Store, with an average rating score of 4.6 out of 5 and with average raters 8,000. Some of these apps were not returned since they do not mention AI or ML in their app description. These 5 apps are Wysa (Beatty et al, 2022), Youper (Mehta et al., 2021), Woebot (Wan, 2021), Elomia (Romanovskyi. et al, 2021), and Happify (Boucher et al., 2021). This is lead to final set of 13 apps, 10 of which are also available on Google Play. These 13 apps belong to two categories: health and fitness (12 apps), and lifestyle (1 app). The analysis consisted of expert evaluation and an auto-ethnography approach. The first author, with expertise in Machine Learning , downloaded and used All the apps extensively for 3-4 weeks on a daily basis. However, Elomia app only used for a duration of 3 days each, as they required a subscription for the premium version. The evaluation of the apps was then conducted by the second author, with expertise in Human-Computer Interaction (HCI). This two-stage evaluation process, combining expertise in ML and HCI, ensured a comprehensive assessment of the mobile app. For expert evaluation, we looked at apps’ descriptions on Apple Store, apps’ websites, apps’ Terms and Conditions, and Privacy Policies on marketplace, and in academic papers describing the apps developed in academia. In these documents, we explored aspects related to ethics such as data privacy, ownership, and those concerning AI or ML (Table 2). This table presents a summary of the inadequate explanation of AI across all the documents and the lack of enough knowledge provided to users viewing AI transparency. It shows up the absence of explainable artificial intelligence (XAI) practices in these documents, which can lead to obstacles in understanding how AI make decisions. Users may not have access to clear explanations or insights into the underlying mechanisms of AI, limiting their ability to comprehend the reasoning behind AI-driven outcomes. Furthermore, we have downloaded and actively used these apps for a period of two weeks to thoroughly review and evaluate their functionalities. This allowed us to gain insights into how the apps work and understand their features in detail. The main functionalities are consistent with previous work (Bowie-DaBreo et al., 2020; Qu et al., 2020; Sanches et al., 2019; Thieme et al., 2020) and include tracking mood and emotion, detecting symptoms, and recommending/providing interventions (Table 2).

4. FINDINGS

This section highlights the limited support for AI literacy and explainability in the reviewed apps, along with ethical concerns regarding safety, privacy, biases, and data reliability. The 13 reviewed apps support functionalities such as mood tracking, detection, and personalised therapy recommendations, including conversational agents for talking therapy.

4.1 Limited Support for AI Literacy and Explainability

Findings indicate surprisingly limited support for AI literacy and XAI. We have seen limited mentioning of AI or ML in apps’ descriptions, privacy documents, or terms of conditions. Thus, the 9 apps that mentioned AI or ML provided scarce details, while the remaining 4 apps did not mention AI or ML in their Apple Store’s descriptions, privacy documents and terms and conditions. Among the former 9 apps, 4 of them (VOS, Magnify Wellness, Elomia, and Reflectly) mentioned AI or ML only on their Apple Store’s descriptions with limited details such as “powered by smart AI”, “driven by Artificial Intelligence”, “use AI, or AI-based mental health”, 2 apps (AI Mood and Replika) mentioned these terms in two of these sources (Table 2 column 1 and 4), and 3 other apps (Anima AI Friend: Chat Bot, Diarly, Wysa) mentioned them in all three sources (Table 2 column 2-4) but with no further details. One app (Enlighten) mentioned these terms on its website, with information that the app learns from the user and checks in over time so that the user’s track is always customised to suit user needs and deliver optimal results such as recommended exercises. Additional details were provided by the Diarly app through an external file on the developer’s website about using AI assistant. Not surprisingly, compared to rest of the apps, the 5 apps mentioned in academic papers (Youper, Woebot, Elomia, Wysa, Rplika) provided richer AI details in these papers (Beatty et al., 2022; Mehta et al., 2021; Possati,

2022; Romanovskyi et al., 2021; Wan, 2021). In contrast, only one other app: Happify was mentioned in scholarly work, however, without reference to AI (Boucher et al., 2021).

4.2 Ethical Consideration

4.2.1 Risks: Children's Safety and Privacy

We now describe ethical issues regarding risk of harm due to AI-based apps and their concerns regarding users' safety and privacy. Findings indicate that while for 5 apps (Wysa, Woebot, Elomia, Magnify Wellness, Replika) the age of use specified in app's description matches the age mentioned in the privacy policy and terms of conditions, the remaining 8 apps provide inconsistent information regarding users' recommended age. From these 8 apps, 5 of them me.ntioned in their description on Apple Store the age of use as 4+ but an older age in privacy policy, namely 13 years (Reflectly and VOS) or 16 years (Happify), or no specific age (Diarly, Online Therapy: AI Mood+ Diary). In addition, three other apps mentioned in their descriptions the age of use is 12+, but in the privacy and policy the age is 18+ (Enlighten, Youper, Anima AI Friend: Chat Bot). These discrepancies raise concerns about children's safety and privacy, especially given their increased vulnerability due to potential mental health conditions.

4.2.2 Risks: Data Reliability

Most of the apps ask users to enter or record their data manually through self-reports of emotions or moods. (Table 1 columns 1-5). Incomplete, underrepresented, or inaccurate data may lead to algorithmic biases. As a result, these can impact the accuracy of the detected emotion or recommended intervention.

4.2.3 Risks: ML-based Biases

Study findings show that NLP models are used in 11 apps. However, such models provide limited account of users' demographics. These models are often trained on large sets of text data that may or may not be representative of all language variants, dialects, or cultures (Colombo et al., 2020). Furthermore, NLP models rely on statistical patterns rather than comprehension of the meaning of language and its context. As a result, they may not always function well with a wide range of linguistic inputs and may provide unexpected or erroneous findings when processing text is shaped by cultural or demographic characteristics. As a result, AI chatbots are prone to selection bias as they learn from users with different relevant characteristics (Abd-Alrazaq et al., 2020). Another illustration of ML-based bias concerns Generative Pre-Trained Transformer-3 (GPT3), a machine-learning model pretrained on large corpus of text through unsupervised learning to generate human-like written language responses (Saha et al., 2019). Both Replika and Diarly app use GPT-3, and it is likely that its use is common in other chatbots. Recently, there has been increasing GPT-3 concerns in regard to unintentional harm caused by gender or age bias (Floridi & Chiriatti, 2020) impacting the accuracy of its output (Saha et al., 2019).

4.3 AI-Based Interventions

This section details the 4 main AI-based interventions for tracking and detection mood or emotion, providing recommendations for indicate the predominant use of AI for providing personalised recommendations for therapy, and supporting talking therapy through conversational agents.

4.3.1 Tracking Emotions and Moods

Findings show that most of the apps support predominantly the tracking of moods or emotion (7 apps), and to a lesser extent the tracking of mental health symptoms (2 apps) or behaviour habits (1 app). In addition, 4 apps can be integrated with the Apple Health apps for tracking exercises and visualising activity reports. Apple Health app stores health and fitness information from one's phone and Apple watch including physical activity and biosensing data. With respect to the specific data modality for capturing moods, emotions, or symptoms, outcomes reveal that most of the reviewed apps use free text manually entered by users (9 apps), or multimodal data with text and photos (1 app), or text, photos, or transcripts of the conversation with the chatbot (voice notes) (3 apps). The tracking functionality per sei does not involve AI techniques or models but generates that data set to be used by such models for detecting the mood or emotion, and for providing recommendations for personalised interventions and/or their delivery as further described.

4.3.2 Detecting Emotions and Moods

Although detection is considered one of the most common functions of AI-based systems, only 2 apps (Online Therapy: AI Mood+ Diary, Replika) used such AI models to detect moods or emotions. Online Therapy: AI Mood+ Diary does so by analysing a photo of user's face and their emotional expressions using sentiment analysis and it outcomes include the identified emotions and their probability such as sad 40%, angry 20%, and fear 40%. This app does not follow up with recommendations for therapeutic interventions. For mood detection, this app asks predefined questions related to mood such as how do you feel? Then it will recommend interventions such as mindfulness-based CBT or exercises. Replika tracks user's mood through the text during the conversation. In contrast with this limited used of AI-based models for detecting affective states, most apps employ structured questionnaires such as Patient Health Questionnaire (PHQ-9) (Kroenke et

al, 2001) (6 apps). The use of this valid and reliable scale by almost half of the apps is a positive outcome. Findings also indicate a missed opportunity regarding the detection of moods of emotions given than many apps (7 apps) track them through text-based data, albeit with limited use of AI/ML techniques. AI techniques such as text classification which can be used to classify emotional text and detect negative emotions which can be then used to provide personalised recommendations or feedback in the context of conversational agents.

4.3.3 Providing Recommendations for Therapy or Wellbeing Interventions

Several apps provide recommendations for CBT or mindfulness-based CBT informed by users' tracked mood, speech, or mental health symptoms of depression or anxiety. Such tracking involves scale PHQ-9 (6 apps), free text (1 app), or both PHQ-9 and free text (1 app). The AI recommendation engine involves techniques such NLP and decision tree to providing such personalised interventions based on the user's mood can also generate graphs showing the correlations of mood and other tracked behaviours such as sleep, supporting users to self-monitoring and self-regulate such behaviours (Table 1 col 9). In particular, the use of NLP in the 3 apps (Woebot, Wysa, Youper) with free text allows for personalised recommendations. For example, the Youper app uses both free-text in conversational interface and PHQ-9 for tracking mood. In the daily checkin, the app asks about emotional states and activity, then the AI chatbot suggests intervention. Other 6 apps use PHQ-9 in addition to NLP to detect emotions or moods and to recognize behaviour patterns in journal entries in order to provide personalised recommendations such as meditation, reading or exercises.

4.3.4 Supporting Talking Therapy Conversational Agents

An important outcome is that 8 of the 13 apps involve conversational agents. 5 of these apps integrate psychological interventions such as CBT or Mindfulness-based CBT, at to a lesser extent those informed by Positive psychology. The chatbots aim to deliver such interventions through conversation with users with mental health conditions particularly depression or anxiety. These 5 apps offer CBT sessions, tailored to individual's mental health needs based on initial questions that users are asked either open questions or structured ones in validated scales for screening for depression such as PHQ-9. The CBT sessions can also be customised for specific groups such as pregnant women, workers, and students (Youper). This app counting steps and calculate sleep's hours. The app also captures texts and used NLP to understand users' responses and interpret the free-text answers (Floridi & Chiriatti, 2020). With respect to chatbot-based CBT interventions, the Youper app uses a decision tree model to interact with the users by selecting the right response based on user input (Mehta et al., 2021). In addition to the 5 apps providing chatbot-based CBT interventions, the remaining 3 apps employing chatbots do not deliver taking therapy interventions but rely on text-based communication to discuss about topics of interest for the user, alleviate loneliness, and support wellbeing. The conversation is supported by added features like playing games with the chatbot such as riddles, mind reading, and trivia, All chatbots are voice-based so they track users' speech using NLP and generate responses using NLP and GPT-3. In particular, 2 apps use only text (Replika, Anima AI Friend: Chat Bot). However, Replika app offers additional features like photos that users can post in the chatbot, and voice notes as written records of the conversations (only in premium version). The chatbot response is personalised based on the information learned about the user. Diarly provides an AI assistant feature (in the premium version) to support user's daily journaling through scaffolding questions. The responses provided by the chatbot are generated using the Generative Pre-trained Transformer 3 (GPT3) neural network language model. Most of the apps provide users the option to rate the chatbots' performance which can be used to further improve the quality of dialogue. An important outcome is that these apps providing chatbots cannot be integrated with other health and wearable devices. We have seen that only 4 of the 13 apps can be integrated with the Apple Health app, and none of the remaining apps leverage biosignal data such as heart rate or electrodermal activity, both important for emotional awareness and regulation (Colombo et al., 2020). To summarise, the limited mention of AI or ML in apps' descriptions provides insufficient support for users' awareness and understanding of of these technologies, and how the apps process data, make decisions, or provide recommendations, which in turn can lower users' trust in such apps.

5. DISCUSSION AND DESIGN IMPLICATIONS

Our study aimed to address the gap of the limited research on AI-based mental health technologies, particularly mobile apps. We now revising the research questions to highlight novel insights including the need for further research on conversational agents for CBT interventions leveraging multimodal data, while addressing the NLP ethical biases, and for user studies to compare AI-based models and the XAI of their outputs.

5.1 Towards Conversational Agents-based CBT Interventions Leveraging Multimodal Data.

With respect to the first one on the main functionalities of the top-rated AI-based mental health apps, findings indicate the predominant use of AI for providing personalised recommendations

for therapy and wellbeing interventions, and in particular the use of voice-based conversational agents for supporting talking therapy such as CBT-based ones. These are important outcomes, contrasting the limited use of personalised interventions in HCI research on mental health which has focused mostly on tracking and diagnosing such conditions (Sanches et al., 2019). Our outcomes also extend findings on previous reviews of mobile apps for depression whose functionalities involve tracking as well as interventions, albeit limited use of personalised CBT-based ones such as thought diaries and without underpinning ML techniques (Qu et al., 2020). Regarding main functionalities, our findings showed that both the provision of recommendations for therapy and wellbeing interventions, as well as the provision of interventions most often through conversational agents is based on tracked data, predominantly text based and less so on AI-based detection of emotions. Another surprising outcome is the limited use of multimodal data, such as physical activity captured by mobile phones or wearables, in AI-based mental health apps. Combining emotional, physical, and contextual data can provide a more comprehensive view of an individual's mental well-being. While some depression apps have incorporated multimodal data to a limited extent (Qu et al., 2020), there is potential for AI-based apps to leverage this data by integrating smartphone and smartwatch apps, since different biodata may increase the accuracy of AI models (Kim et al., 2022)

5.2 Towards Clinically Valid Conversational Agents Addressing the Ethical Concerns of NLP Biases

For the second research question we looked at the ethical challenges of AI-based mental health apps. Our outcomes confirm such challenges (Zajac et al., 2023; Kapania et al., 2022) and also identified three main sources of harms: (i) the privacy of personal data of the vulnerable users living with mental health conditions and particularly children, (ii) the biases of ML models and algorithms, and (iii) the reliability of tracked data. These can be addressed through stronger guidelines on Apple Store regarding consistent information for the appropriate age of using these apps. Issues regarding data reliability can also be addressed through complementary multimodal data where user entered data is extended with automatically tracked data through phone or wearables which can also provide much needed larger data sets (Zajac et al., 2023). For app developers, we suggest enabling multi-factor authentication (Alanazi & Aborokbah, 2022) to ensure that children only use such age-appropriate mental apps under adult supervision (Lewis, 2020). Similar concerns regarding age have been identified with respect to apps for depression (Sanches et al., 2019) but in the case of AI-based apps for mental health the ethical implications are arguably more concerning, and much care is needed to safeguard children' safety and privacy. In the light of these findings, we also emphasise the need to carefully tailor the AI-based approaches to account for children's related AI biases (Cheng et al., 2019) their specific symptoms and experiences of mental ill health, and their increased challenges of understanding AI-based apps (Druga et al., 2022) It is important to develop a specific approach for privacy and explainable AI for children and narrow down the range of AI use based on age group (Floridi & Chiriatti, 2020; Qu et al., 2020). The algorithmic biases pertaining particularly to the extensively used NLP models in our reviewed apps require attention, given their discrimination and failure to account for demographic and cultural differences (Harrington et al., 2022). Since previous work has shown limited use of NLP models in affective health technologies (Sanches et al., 2019) our findings open up design and research opportunities for exploring the ethical co-design of voice based conversational chatbots while creatively addressing the ethical challenges of NLP models (Grové, 2021). The validity of digitally delivered therapeutic interventions remains problematic especially with regard to mobile apps (Bowie-DaBreo et al., 2020; Khazaal et al., 2018; Søgaard et al., 2019) and more work is needed in this direction. Our outcomes open up research and design opportunities to better support voice-based conversational agents for delivering talking therapy interventions such as CBT, in order to strengthen their clinical validity and address their outstanding ethical concerns regarding users' personal data (Bowie-DaBreo et al., 2020; Khazaal et al., 2018).

5.3 Towards User's Comparative Exploration of AI-based Models and their Outputs' XAI

Our final research question focused on how the AI-based apps for mental health support users' understanding of their AI models and outputs. Study outcomes highlight a surprisingly low support for AI literacy and XAI, with limited information on AI models and even less on their XAI despite our exploration of several sources such as apps' privacy documents, terms of conditions, or description on Apple Store. One way to address this is through explicit and clear information that Apple Store developers are required to provide for their AI-based apps. At least, such information should describe the data set in terms of content, i.e., emotions, thoughts, modality i.e. text, photos, and mode of capture: user entered, or automatically recorded, the AI models, i.e., supervised, unsupervised learning, and specific models or techniques, AI outputs: which models are used on what data sets to generate specific outputs, and relevant algorithmic biases, their impact on outputs and developers' effort for mitigating them. Given the vulnerable users of these apps, these findings on limited AI literacy and XAI are

concerning. Such users may need additional support since understanding and trusting AI-based technologies more broadly is particularly challenging (Chalabianloo et al., 2022; Bowie-DaBreo et al., 2020; Warren et al., 2022). We suggest applying model transparency (Sampson et al., 2019) to address algorithmic bias through clear insight of the decisions making process or how the app arrived at specific recommendations (Saha et al., 2019). We also suggest the use of fairness metrics (Garg et al., 2020) by comparing the outcomes of an algorithm for different groups of people, based on demographic factors such as race, gender, and age. Towards supporting XAI, emerging work has focused on studies comparing a range of AI models or techniques and the XAI of their outcomes (Kapania et al., 2022; Sanches et al., 2019) such as those for recommender systems (Langer et al., 2022) but their application to mental health domain remains limited (Chalabianloo et al., 2022; Kim et al., 2022). We suggest the use of SHAP (Lundberg & Lee, 2017), an open source, game-based approach for explaining the outcomes of ML models to support the comparison of their XAI, recently used on multimodal biosensing data for understating stress detection (Chalabianloo et al., 2022). In addition, XAI approaches can be particularly used to explain why an NLP model made a specific prediction, allowing biases in the model's decision-making process to be identified and corrected (Danilevsky.,2021).

6. CONCLUSION

We report functionality review of 13 AI-based mental health apps from Apple Store. Study outcomes suggest limited AI literacy and explainability and four main functionalities of these apps namely tracking and detecting emotions and moods, providing recommendations for therapy and wellbeing interventions, and supporting talking therapy through conversational agents powered by NLP models. We discuss these outcomes and articulate design implications for developing conversational agents to support CBT interventions based on tracked multimodal data, addressing the ethics of NLP biases, and of user exploration of AI-based models and their XAI.

7. REFERENCES

Abd-Alrazaq, A.A., Rababeh, A., Alajlani, M., Bewick, B.M. and Househ, M (2020) Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. Journal of medical Internet research, 22(7), p.e16021.

Alanazi, M. and Aborokbah, M (2022) Multifactor Authentication Approach on Internet of Things: Children's Toys. In 2022 2nd International Conference on Computing and Information Technology (ICCIT) (pp. 6-9). IEEE.

Alfaras, M., Tsaknaki, V., Sanches, P., Windlin, C.,Umair, M., Sas, C. and Höök, K (2020) Frombiodata to somadata. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-14).

Balkir, E., Kiritchenko, S., Nejadgholi, I. and Fraser,K.C (2022) Challenges in applying explainability methods to improve the fairness of NLP models. arXiv preprint arXiv:2206.03945.

Beatty, C., Malik, T., Meheli, S. and Sinha, C (2022) Evaluating the Therapeutic Alliance with a Free-

Text CBT Conversational Agent (Wysa): A MixedMethods Study. Frontiers in Digital Health, 4, p.847991.

Boucher, E.M., McNaughton, E.C., Harake, N., Stafford, J.L. and Parks, A.C (2021). The impact of a digital intervention (Happify) on loneliness during COVID19: qualitative focus group. JMIR mental health, 8(2), p.e26617.

Bowie-DaBreo, D., Sünram-Lea, S.I., Sas, C. and Iles-Smith, H., 2020. Evaluation of treatment descriptions and alignment with clinical guidance of apps for depression on app stores: systematic search and content analysis. JMIR formative research, 4(11), p.e14988.

Chalabianloo, N., Can, Y.S., Umair, M., Sas, C. and Ersoy, C (2022) Application level performance evaluation of wearable devices for stress classification with explainable AI. Pervasive and Mobile Computing, 87, p.101703.

Cheng, H.F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F.M. and Zhu, H (2019) Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In Proceedings of the 2019 CHI conference on human factors in computing systems (pp. 1-12).

Cila, N (2022) Designing Human-Agent Collaborations: Commitment, responsiveness, and support. In CHI Conference on Human Factors in Computing Systems (pp. 1-18).

Colombo, D., Fernández-Álvarez, J., Suso-Ribera, C., Cipresso, P., Valev, H., Leufkens, T., Sas, C., Garcia-Palacios, A., Riva, G. and Botella, C (2020) The need for change: Understanding emotion regulation antecedents and consequences using ecological momentary assessment. Emotion, 20(1), p.30.

Danilevsky, M., Dhanorkar, S., Li, Y., Popa, L., Qian, K. and Xu, A (2021) Explainability for natural language processing. In Proceedings of the 27th ACM SIGKDD *Conference on Knowledge Discovery & Data Mining* (pp. 4033-4034).

Druga, S., Christoph, F.L. and Ko, A.J (2022). Family as a Third Space for AI Literacies: How do children and parents learn about AI together*?. In CHI Conference on Human Factors in Computing Systems* (pp. 1-17).

Ehsan, U., Liao, Q.V., Muller, M., Riedl, M.O. and Weisz, J.D (2021) Expanding explainability: Towards social transparency in ai systems*. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-19).

Ehsan, U., Tambwekar, P., Chan, L., Harrison, B. and Riedl, M.O (2019) March. Automated rationale generation: a technique for explainable AI and its effects on human perceptions*. In Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 263-274).

Floridi, L. and Chiriatti, M (2020) GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4), pp.681-694.

Garg, P., Villasenor, J. and Foggo, V (2020) December. Fairness metrics: A comparative analysis. *In 2020 IEEE International Conference on Big Data* (Big Data) (pp. 3662-3666). IEEE.

Grové, Christine. "Co-developing a mental health and wellbeing chatbot with and for young people." Frontiers in psychiatry 11 (2021): 606041.

Harrington, C.N., Garg, R., Woodward, A. and Williams, D (2022) "It's Kind of Like Code-Switching": Black Older Adults' Experiences with a Voice Assistant for Health Information Seeking. *In CHI Conference on Human Factors in Computing Systems* (pp. 1-15).

Jain, V., Agarwal, P (2017). Symptomatic diagnosis and prognosis of psychiatric disorders through personal gadgets. *In Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 118-123).

Kapania, S., Siy, O., Clapper, G., SP, A.M. and Sambasivan, N (2022)" Because AI is 100% right and safe": User Attitudes and Sources of AI Authority in India. *In CHI Conference on Human Factors in Computing Systems* (pp. 1-18).

Khazaal, Y., Favrod, J., Sort, A., Borgeat, F. and Bouchard, S (2018) Computers and games for mental health and well-being. *Frontiers in psychiatry*, 9, p.141.

Kim, T., Kim, H., Lee, H.Y., Goh, H., Abdigapporov, S., Jeong, M., Cho, H., Han, K., Noh, Y., Lee, S.J. and Hong, H (2022) Prediction for Retrospection: Integrating Algorithmic Stress Prediction into Personal Informatics Systems for College Students' Mental Health. *In CHI Conference on Human Factors in Computing Systems* (pp. 1-20).

Kroenke, Kurt, Robert L. Spitzer, and Janet BW Williams. "The PHQ-9: validity of a brief depression severity measure." *Journal of general internal medicine*16, no. 9 (2001): 606-613

Langer, M., Hunsicker, T., Feldkamp, T., König, C.J. and Grgić-Hlača, N (2022) April. "Look! It'sa Computer Program! It's an Algorithm! It's AI!": DoesTerminology Affect Human Perceptions and Evaluations of Algorithmic Decision-Making Systems?. *In CHI Conference on Human Factors in Computing Systems* (pp. 1-28).

Lewis, N. (2020) Design and development of a soft skills acquisition application for young children in informal contexts (Doctoral dissertation, Cape Peninsula University of Technology).

Liao, M., Sundar, S.S. and B. Walther, J (2022) User Trust in Recommendation Systems: A comparison of Content-Based, Collaborative and Demographic Filtering. *In CHI Conference on Human Factors in Computing Systems* (pp. 1-14).

Loi, D., Wolf, C.T., Blomberg, J.L., Arar, R. and Brereton, M (2019) Co-designing AI futures: Integrating AI ethics, social computing, and design. *In Companion publication of the 2019 on designing interactive systems conference 2019 companion* (pp. 381-384).

Long, D. and Magerko, B (2020) What is AI literacy? Competencies and design considerations. In Proceedings of the 2020 CHI conference on human factors in computing systems (pp. 1-16).

Lundberg, Scott M., and Su-In Lee (2017) A unified approach to interpreting model predictions." *Advances in neural information processing systems 30*.

Mehta, A., Niles, A.N., Vargas, J.H., Marafon, T., Couto, D.D. and Gross, J.J (2021) Acceptability and Effectiveness of Artificial Intelligence Therapy for Anxiety and Depression (Youper): Longitudinal Observational Study. *Journal of Medical Internet Research*, 23(6), p.e26771

Miri, P., Flory, R., Uusberg, A., Culbertson, H., Harvey, R.H., Kelman, A., Peper, D.E., Gross, J.J., Isbister, K. and Marzullo, K (2020) PIV: Placement, pattern, and personalization of an inconspicuous vibrotactile breathing pacer. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 27(1), pp.1-44.

Mohseni, S., Zarei, N. and Ragan, E.D (2021) A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems* (TiiS), 11(3-4), pp.1-45.

Nikiforos, S., Tzanavaris, S. and Kermanidis, K.L., (2020) Virtual learning communities (VLCs) rethinking: influence on behavior modification—bullying detection through machine learning and natural language processing. *Journal of Computers in Education*, 7, pp.531-551.

Possati, L.M., 2022. Psychoanalyzing artificial intelligence: the case of Replika. AI & SOCIETY, pp.1-14.

Qu, C., Sas, C., Roquet, C.D. and Doherty, G., 2020. Functionality of top-rated mobile apps for depression: systematic search and evaluation. *JMIR mental health*, 7(1), p.e15321.

Razak, F.H.A., Hafit, H., Sedi, N., Zubaidi, N.A. and Haron, H., 2010, December. Usability testing with children: Laboratory vs field studies. I*nternational Conference on User Science and Engineering* (i-USEr) (pp. 104-109). IEEE.

Richards, D., Enrique, A., Palacios, J., Eilert, N., Duffy, D., Doherty, G., Jardine, J., Vigano, N. and Tierney, K (2023) SilverCloud Health: Online Mental Health and Wellbeing Platform. *In Digital Therapeutics* (pp. 307-330). Chapman and Hall/CRC.

Rojat, T., Puget, R., Filliat, D., Del Ser, J., Gelin, R. and Díaz-Rodríguez, N (2021) Explainable artificial intelligence (xai) on timeseries data: A survey. arXiv preprint arXiv:2104.00950.

Roli Khanna, Jonathan Dodge, Andrew Anderson, Rupika Dikkala, Jed Irvine, Zeyad Shureih, Kin-Ho Lam, Caleb R. Matthews, Zhengxian Lin, Minsuk Kahng, Alan Fern, and Margaret Burnett. (2022) Finding AI's Faults with AAR/AI: An Empirical Study. ACM Trans. *Interact. Intell. Syst.* 12, 1, Article 1 (March 2022), 33 pages. https://doi.org/10.1145/3487065

Romanovskyi, O., Pidbutska, N. and Knysh, A., (2021) Elomia Chatbot: The Effectiveness of Artificial Intelligence in the Fight for Mental Health. In COLINS (pp. 1215-1224).'

Saha, K., Kim, S.C., Reddy, M.D., Carter, A.J., Sharma, E., Haimson, O.L. and De Choudhury, M (2019) The language of LGBTQ+ minority stress experiences on social media. *Proceedings of the ACM on human-computer interaction*, 3(CSCW), pp.1-22.

Sampson, C.J., Arnold, R., Bryan, S., Clarke, P., Ekins, S., Hatswell, A., Hawkins, N., Langham, S., Marshall, D., Sadatsafavi, M. and Sullivan, W (2019) Transparency in decision modelling: what, why, who and how?. Pharmacoeconomics, 37(11), pp.1355-1369.

Sanches, P., Janson, A., Karpashevich, P., Nadal, C., Qu, C., Daudén Roquet, C., Umair, M., Windlin, C., Doherty, G., Höök, K. and Sas, C., (2019) HCI and Affective Health: Taking stock of a decade of studies and charting future research directions. *In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-17).

Simard, F., Kizuk, S.A. and Fortin, P.E (2021) Mindscape: Transforming Multimodal Physiological Signals into an Application Specific Reference Frame. In *Companion Publication of the 2021 International Conference on Multimodal Interaction* (pp. 334-336).

Søgaard Neilsen, A. and Wilson, R.L (2019) Combining e-mental health intervention development with human computer interaction (HCI) design to enhance technology-facilitated

Straw, I. and Callison-Burch, C., 2020. Artificial Intelligence in mental health and the biases of language-based models. PloS one, 15(12), p.e0240376.

Terzimehić, N., Häuslschmid, R., Hussmann, H. and Schraefel, M.C (2019) A review & analysis of mindfulness research in HCI: Framing current lines of research and future opportunities. *In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-13).

Thieme, A., Belgrave, D. and Doherty, G (2020) Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM Transactions on Computer-Human Interaction* (TOCHI), 27(5), pp.1-53.

Thieme, A., Hanratty, M., Lyons, M., Palacios, J.E., Marques, R., Morrison, C. and Doherty, G (2022) Designing Human-Centered AI for Mental Health: Developing Clinically Relevant Applications for Online CBT Treatment. *ACM Transactions on Computer-Human Interaction.Transactions on Interactive Intelligent Systems* (TiiS), 11(3-4), pp.1-45. https://doi.org/10.1007/s40692-020-00166-5Umair, M., Chalabianloo, N., Sas, C. and Ersoy, C., (2021) HRV and stress: A mixed-methods approach for comparison of wearable heart rate sensors for biofeedback. *IEEE Access*, 9, pp.14005-14024.

Wan, E (2021) " I'm like a wise little person": Notes on the Metal Performance of Woebot the Mental Health Chatbot. *Theatre Journal*, 73(3), pp.E-21.

Warren, G., Keane, M.T. and Byrne, R.M (2022) Features of Explainability: How users understand counterfactual and causal explanations for categorical and continuous features in XAI. arXiv preprint arXiv:2204.10152.

Zając, H.D., Li, D., Dai, X., Carlsen, J.F., Kensing,F. and Andersen, T.O (2023) Clinician-facing AI in the Wild: Taking Stock of the Sociotechnical Challenges and Opportunities for HCI. ACM Transactions on Computer-Human Interaction, 30(2), pp.1-39.

***Table 1**. Apps and their main functionalities*

App name	Mental health condition	Therapy intervention	AI-based intervention	User entered data	Scale	Tracking	Detecting	AI techniques	Outcomes of AI-based intervention
Anima AI Friend	Generic	No CBT	Conversational agent	Text	Free text	Speech		NLP	To understand and respond to users' emotional states To provide personalised emotional support through conversations
Diarly	Wellbeing	No CBT	Self-Monitoring Conversational agent	Text	Free text	Mood		NLP Sentiment analysis Text classification GPT-3	To provide personalised emotional support during user's journaling
Elomia	Generic	CBT	Conversational agent Recommendation	Text	Free text	Speech		NER COSMIC DIALOGPT SQuAD GECToR Text classification	To identify emotions, names, locations To answer user' questions To provide recommendations for exercises to reduce anxiety
Enlighten	Generic Wellbeing	CBT DBT Mindfulness-CBT	Recommendation	Text	PHQ-9	Mood Habit	Symptoms	NLP Sentiment analysis Text classification	To provide personalised recommendations for meditation, relaxation, focus, and sleep based on emotion
Happify	Wellbeing	CBT Mindfulness-CBT Positive psychology	Recommendation	Text	PHQ-9	Mood		NLP Text classification	To recognise behaviour patterns-based users input, and emotion in users' journal To provide personalised recommendations for self-care activities
Magnify Wellness	Generic	Positive psychology	Conversational agent Recommendation	Text	PHQ-9	Mood		NLP	To provide personalised recommendations for self-care and exercises
Online Therapy: AI Mood+ Diary	Anxiety Stress	CBT	Recommendation	Text Photos	PHQ-9	Mood	Emotion	NLP CNN Text classification	To recognise emotion from user's face To recommend meditation and motivation based on tracked mood
Reflectly	Wellbeing	Mindfulness CBT	Self-Monitoring	Text Photos Voice notes	PHQ-9	Mood		NLP	To recognise emotion based on user's journal and personalised daily challenge. To show mood chart based on time period to support self-monitoring
Replika	Anxiety Depression Wellbeing	No CBT	Conversational agent	Text Photos Voice notes	Free text	Speech		ANNs GPT-3	To model human behaviour and language patterns
VOS: Well-being Plan & Journal	Generic Wellbeing	CBT	Recommendation	Text	PHQ-9	Mood		NLP	To analyse motion patterns in journal entries To provide personalised recommendations To track mood correlations with sleep and kcal
Woebot	Generic	DBT Mindfulness-CBT	Conversational agent	Text	Free text	Mood		NLP Text classification	To identify emotions based on sentiment analysis To provide personalised recommendations
Wysa	Generic Wellbeing	CBT DBT	Conversational agent Recommendation	Text	Free text	Anxiety or depression symptoms	Mood	NLP	To identify emotions based on sentiment analysis To provide personalised recommendations
Youper	Anxiety Depression	CBT	Conversational agent Recommendation	Text Photos	Free text PHQ-9	Mood Symptoms	Emotion	NLP Deep learning Decision tree	To identify emotions based on sentiment analysis To provide personalised recommendations

***Table2**: AI Explainability of the mental health apps*

App Name	Apple store description	Terms of conditions	Privacy policy	Other
Anima AI Friend	Yes. No details	Yes. No details	Yes. No details	No
Diarly	Yes. No details	No	No	Yes. No details
Elomia	Yes. No details	No	No	Yes
Enlighten	No	No	No	Yes. No details
Happify	No	No	No	Yes. No details
Magnify Wellness	Yes. No details	No	No	No
Online Therapy: AI Mood+ Diary	Yes. No details	No	Yes. No details	No
Reflectly	Yes. No details	No	No	No
Replika	Yes. No details	No	Yes. No details	Yes
VOS: Well-being Plan & Journal	Yes. No details	No	No	No
Woebot	No	No	No	Yes
Wysa	Yes. No details	Yes. No details	Yes. No details	Yes
Youper	No	No	No	Yes

***Figure 1:** PRISMA diagram of the apps' selection process*