Mental Health Chatbots and AI-Administered Treatment in Psychotherapy: A Systematic Literature Review

Kristina A. Algas

University of California, Santa Barbara

WRIT 109ST: Writing for Science and Technology

Prof. Kenneth Smith

May 22, 2020

Introduction

The World Health Organization (WHO) reported in 2018 that approximately 1 in 10 individuals worldwide are in need of mental healthcare treatment (World Health Organization, 2018). Especially in recent years, growing prevalence and public awareness of mental health disorders has led to a surge in demand for mental healthcare (American Psychology Association, 2021). However, previous studies suggest that not everyone who has access to mental healthcare receives adequate treatment: Swift & Greenberg (2012) found that approximately 19.7% of all psychotherapy patients choose to withdraw early from treatment. Lindarnon et al. (2018) observed a similar percentage of dropout rates within clinical populations suffering from symptoms of depression (20.6%) and anxiety (16.5%). Moreover, early withdrawal from psychotherapy treatment has been identified as a predictor for overall poorer mental health outcomes post-treatment (Barrett et al., 2008). As such, strategies for improving the quality and efficacy of psychotherapy treatment have been the subject of a large body of existing research.

The rapid evolution of artificial intelligence (AI) in recent years has presented an alternative, novel solution for increasing psychotherapy retention rates. The adaptation of conversational AI programs, or *chatbots*, has been tentatively investigated as a potential supplement-- if not a full replacement-- for human-administered psychotherapy treatments. *Chatbots* are autonomous computer programs that can learn from and synchronously respond to user input. Researchers in the field of *human-computer communication* (HCC) have consistently supported the notion that humans are capable of perceiving, interacting, and empathizing with chatbots as comparable-- albeit not equivalent to-- autonomous living beings (Hortensius et al., 2018; Mishra, 2006). The findings of previous studies, combined with the low maintenance cost of AI programs, show that AI-administered treatment (AAT, for the purposes of the present study) technologies have no shortage of potential in the field of psychotherapeutic treatment.

Recent developments in autonomous AI technology have brought about the creation of chatbots specifically designed to help maintain users' mental health (e.g. WYSA, Woebot, etc.). Because this particular branch of AI technology development is highly novel, even within the nascent field of HCC, currently available research on the subject of chatbots in mental health is considerably limited; moreover, no published systematic literature reviews on the subject have examined the feasibility of AAT in psychotherapy under the theoretical lens of the Computers are Social Actors (CASA) paradigm, which focuses on the capability of human social actors to interact and form relationships with virtual agents in the same capacity they would with other human beings (Nass et al., 1994). As such, the objective of the present study is to review and summarize the currently available literature on chatbots in psychotherapy in terms of relational development and treatment efficacy, which will thereafter be analyzed in the context of the CASA paradigm.

Methodology

In alignment with the objectives of the review, the following research questions are addressed:

RQ1: Which clinical populations of mental health disorders have been most widely investigated as viable for AI-administered psychotherapy treatment?

RQ2: What are some of the strengths and limitations of AI-administered psychotherapy treatment in comparison to human-administered psychotherapy treatment?

RQ3: What comparisons, if any, can be drawn between the development of AI-human TCR and that of human-human TCR?

Inclusion and Exclusion Criteria

Relevant studies were deemed eligible for review by the following inclusion criteria:

- Studies in which the chatbot is fully autonomous (i.e. capable of operation without manual human input)
- Studies in which eligible participants:
    - Report experiencing ongoing symptoms of at least one mental health disorder prior to the study
    - Report experiencing symptoms of at least one mental health disorder as a symptom of a pre-existing health condition (i.e. physiological conditions, substance abuse, etc.)

Studies were removed from the review if they fell under the following exclusion criteria:

- Secondary studies (i.e. literature reviews, mapping studies)
- Technical documents describing proposed experiments or software developments
- Studies published earlier than Jan. 1, 2017
- Studies in which the software or interface design of mental health chatbots take precedence over the overall efficacy/outcomes of treatment

Key Terms

Articles were gathered for the purposes of this review in May 2022 via a manual database search. The database search was conducted on PubMed using varied combinations of the following keywords: *chatbot*, *mental health*, *therapy*, and *psychotherapy*. Additionally, relevant studies were organized into one or more of the following categories (see Fig. 1).

Focus Dimensions	Relational Development	Focused on the progress and development of the therapeutic alliance throughout AAT
	Efficacy	Focused on the effectiveness of AAT as a novel form of treatment for mental health symptoms
	Comparison	Focused on comparing the effectiveness of AAT to a control group
Clinical Populations	Anxiety Disorders	e.g. General Anxiety DIsorder (classified as "anxiety" if not specified)
	Mood Disorders	e.g. Major Depressive Disorder (classified as "mood disorder" if not specified"
	Developmental Disorders	e.g. Attention Deficit Hyperactivity Disorder (ADHD)
	Substance Abuse	e.g. smoking, substance abuse, alcohol abuse, etc.
	Other	Any problem unrelated to mental health (e.g. physical pain)
	Not Specified	No specific mental health disorder is mentioned

Fig. 1: Study classification categories.

Study Selection

Relevant studies were gathered and identified via a manual database search on PubMed. Applying search terms containing the aforementioned keywords resulted in a total of 26 articles. The titles and abstracts of each article were manually screened, and were excluded from the review if they a) failed to meet all inclusion criteria, or b) met one or more of the exclusion criteria. After applying the inclusion criteria and removing duplicate studies, a total of 7 relevant research articles were approved for review.

Results

Findings

The application of artificial intelligence in psychotherapy treatment is a rapidly growing, albeit highly novel, field of research and development. As such, all 7 articles included in this review were published no earlier than 2018, rendering the exclusion criteria of the present study obsolete in regards to publishing date.

All studies focused on the feasibility and preliminary efficacy of chatbots in treating adults suffering from self-diagnosed mental health disorders (for information on the classification and outcomes of relevant studies, refer to Fig. 2). Overall, a positive correlation between chatbot mental health intervention and overall efficacy of treatment (in terms of TCR quality and/or symptomatic behavior reduction) was observed across all studies. Mental and physical health improvements were found across all studies to be highest in participants who received the most frequent intervention from mental health chatbots.

In studies that included a comparison between a control group (i.e. received no intervention or an alternative method of intervention) and a group that received chatbot-facilitated mental health intervention (Fitzpatrick et al., 2017; Jang et al., 2022; Leo et al., 2022), participants in the latter condition reported a higher improvement in mental health compared to those in the former. Notably, Leo et al. (2022) reported that orthopedic patients who received mental health intervention from a chatbot (WYSA) were more likely to experience reduced chronic physical pain (i.e. pain as a symptom of pre-existing physiological conditions) alongside reductions in symptoms of depression and anxiety in comparison to patients who only received physiological orthopedic treatment.

Study	Sample Size	Chatbot	Clinical Population(s)	Focus Dimension(s)	Description
Beatty et al. (2022)	1,025	WYSA	Anxiety, Major Depressive Disorder	Relational Development	Consistent improvement in therapeutic alliance was observed over the intervention period; participants described and interacted with the chatbot as if it were human, despite being aware of its limitations.
Fitzpatrick et al. (2017)	70	Woebot	Anxiety, Major Depressive Disorder	Efficacy, Comparison (vs. text-only)	Users with high levels of engagement reported the highest average reduction in anxiety and depression symptoms. Participants reportedly attributed perceived emotions, empathy, and anthropomorphic descriptors (e.g. "a fun little dude") to Woebot.
He et al. (2022)	182	flow.ai	Substance Abuse	Relational Development, Efficacy	Participants reported increased motivation to quit after interacting with the chatbot. No significant difference in improvement was observed in the group that received treatment from a chatbot with a more human-like conversation style.
Inkster et al. (2018)	129	WYSA	Major Depressive Disorder	Efficacy	High-engagement users reported a higher average reduction in depression symptoms than low-engagement users.
Jang et al. (2020)	46	Todaki	Developmental Disorder (ADHD)	Efficacy, Comparison (vs. text-only)	Reduced symptoms were only observed in the group that interacted with the chatbot. Participants reacted positively to the chatbot's empathetic disposition, but negatively towards its unnatural conversation style.
Leo et al. (2022)	153	WYSA	Anxiety, Major Depressive Disorder, Other (Orthopedic Conditions)	Comparison (vs. orthopedic treatment only)	The group that received mental health intervention from a chatbot reported reduced symptoms of anxiety, mental health, and physical pain.
Prochaska et al. (2021)	101	Woebot	Substance Abuse	Relational Development, Efficacy	Therapeutic alliance quality and overall willingness to engage in treatment were both positively correlated with reduced instances of substance abuse.

Fig. 2: Descriptions of Relevant Studies

Participants

The quantitative sample size of the participant pool across all studies was 1,709. The lowest minimum age for eligible participants in any given study was n=18, and the highest maximum age was n=65 (Prochaska et al., 2021). Inconsistent reporting in the gender demographics of each study resulted in an inability to determine the gender demographics of each study's participants. Notably, no demographic information was collected by Beatty et al. (2022) due to anonymity concerns; as such, the age and gender demographics of the study's participants (n=1,205) could not be reliably determined.

Each study was categorized by one or more clinical demographics that characterized their participant sample (see Fig. 1). For the purposes of this study, relevant clinical populations were classified as per the classification schema of the DSM-IV. A significant portion of the relevant literature (n=4) focused on the treatment of self-reported symptoms of anxiety and depression; additionally, all studies within this category only required that the participant presented a self-reported diagnosis via one or more diagnostic surveys administered as part of the eligibility screening process. Substance abuse disorders comprised the target clinical populations of (n=2) studies, and developmental disorders (i.e. attention deficit hyperactivity disorder/ADHD) were the focus of one study (Jang et al., 2021).

Software

All relevant studies involved mental health intervention from a mobile app chatbot as part of the experimental procedure. Of these studies, (n=3) investigated the feasibility of WYSA, a mobile app chatbot for mental health, as a supplementary treatment for self-reported symptoms of anxiety disorders and major depressive disorder. (n=2) studies investigated Woebot, a mobile-based chatbot with design capabilities comparable to those of WYSA, as a supplementary treatment for substance abuse behaviors (n=1) and symptoms of anxiety and depression (n=1). Of the remaining studies, (n=1) study used flow.ai, a chatbot specifically designed for the purpose of carrying out artificial intelligence research; and (n=1) study (Jang et al., 2021) involved the development of an entirely new chatbot software (Todaki) for treating behavioral symptoms of ADHD.

In terms of the strengths and shortcomings of currently available technology, limitations such as underdeveloped conversational capabilities and insufficient software formatting were cited as negative traits by participants in studies that gathered qualitative feedback data (Beatty et al., 2021; Fitzpatrick et al., 2017; Inkster et al., 2018; Jang et al., 2021). Additionally, across all studies, voluntary user engagement was observed to decrease significantly across the given length of intervention and observation.

Relevant Study Limitations

There were a number of consistent limitations throughout the relevant studies. Firstly, studies with intervention periods that took place during the beginning of the COVID-19 pandemic (i.e. ca. 2020 - 2021) were conducted remotely via computer-mediated communication methods (i.e. text-based message exchange). As a result, these studies were limited by an inability to consistently monitor client conditions outside of self-reported updates. Additionally, all studies were conducted via limited interventions over the course of two to four weeks, which were consistently reported to be too short of a time frame to determine the long-term efficacy of treatment. Future studies should take place over the course of an extended intervention period in order to obtain more reliable results. Moreover, the continuous trend of user interest dropping in the middle of the intervention period should be addressed in further research and development.

Discussion

Principal Findings

To summarize the findings of the present study, participants across all relevant studies were reported to experience reduced symptoms when interacting with conversational agents, especially when TCR quality was taken into account. Previous studies have found that patients who have a higher quality TCR with their treating therapist have been found to have a greater likelihood of self-disclosure, which thereby results in more effective counseling and a decreased likelihood of early withdrawal (Primavera et al., 2010). The capability of psychotherapists to establish and improve the TCR depend highly on their capabilities for building rapport and engaging in empathic nonverbal behaviors (i.e. taking turns speaking, open body language, etc.) (Foley & Gentile, 2010; Ramseyer & Tschacher, 2011). It follows, then, that adapting these capabilities into mental health chatbots would be instrumental in terms of ensuring treatment efficacy and user engagement.

To this end, participants that received chatbot mental health intervention across all relevant studies were prompted to provide feedback on the chatbot's perceived empathetic capabilities. Positive relational traits (friendliness, empathy, etc.) were consistently reported to be perceived as positive characteristics. Qualitative data from the relevant studies suggests that participants fully perceived the chatbot as autonomous and even friendly (Beatty et al., 2021; Fitzpatrick et al., 2017; Inkster et al., 2018; Jang et al., 2021); an example of this can be observed in the article by Fitzpatrick et al. (2017), in which participants reportedly referred to Woebot with "he" pronouns and ascribed endearing traits to its personality (e.g. "a fun little dude") (p.8)

These findings are consistent with previous literature on empathetic human-chatbot communication, which suggest that humans are capable of empathizing with and forming emotional attachments to sufficiently advanced AI agents. However, chatbots with underdeveloped anthropomorphic traits (e.g. insufficient rapport-building capabilities) have been found to be less capable in terms of improving the TCR with users: for instance, Jang et al. (2022) reported that their newly developed mental health chatbot, Todaki, received negative feedback for having unnatural dialogue patterns and sentence structure. Such conclusions are consistent with previous existing research: an earlier study by Bickmore et al. (2010) on empathic touch in embodied virtual agents reported that participants felt uncomfortable interacting with a chatbot that was unable to convincingly replicate human behavioral patterns, although they were capable of understanding what emotions the agent was attempting to convey.

It is worth noting that participants who interacted with chatbots throughout each study consistently reported a higher willingness to self-disclose. These findings are consistent with existing literature on HCC, which suggests that people are encouraged to exhibit a greater willingness to disclose vulnerable information when interacting with virtual agents (Weisband & Kiesler, 1996; Barrett et al., 2008). This condition only applies when participants believe they are interacting with an autonomous virtual agent, regardless of whether or not said virtual agent is actually autonomous or human-operated (Barrett et al., 2008). One possible explanation for this behavioral pattern could be that, due to the known limitations of current artificial intelligence technology, humans may perceive artificial intelligence agents to be incapable of meaningful judgment or feedback (Mishra, 2009). Despite this, participants across relevant studies reportedly acknowledged and accepted the technological limitations of the chatbots throughout their interactions (Beatty et al., 2022).

The implication that human users are capable of forming and acknowledging the TCR while interacting with chatbots also raises several questions on the perception of virtual agents as social actors. The findings of the present study may appear to intuitively align with the principles of the Computers are Social Actors (CASA) paradigm (Nass et al., 1994); for instance, qualitative feedback from the study by Fitzpatrick et al. (2017) suggests that participants viewed Woebot, a virtual agent, as a gendered social actor with personality traits and adherence to social norms. Moreover, participants across all studies responded socially to the chatbots they interacted with. However, the findings of the present study acknowledge the limitations of this paradigm: in particular, that humans are more likely to perceive virtual agents as lesser social actors who are incapable of making meaningful judgements on human behavior (Mishra et al., 2009). As such, the development of the TCR in AAT cannot be reliably compared to TCR development in human-human psychotherapy treatment, the latter of which inherently involves the acknowledgement of the therapist as a social actor. It could be inferred, then, that the perceived incapability of virtual agents to judge and criticize human behaviors is what makes them such appealing confidants, especially for mental health patients who may fear the social stigma surrounding mental health.

Limitations of the Present Study

Due to the limited availability of time and material resources, the present study is limited primarily by the breadth of the reviewed literature. To elaborate, the articles included in this study were obtained from a single database (PubMed) and manually screened for relevance; as such, potentially relevant articles from other sources may have been excluded from the present study. In order to ensure the inclusion of all potentially relevant studies, future reviews on the existing literature should be carried out by a more organized research effort with access to wider resources and a more flexible time frame.

Implications for Future Research

The development of consumer-grade, autonomous virtual agents has progressed rapidly in recent years, and is expected to continue further into the future. Novel developments in the field show promise regarding the feasibility of using conversational chatbots to administer remote psychotherapeutic treatment to clinical populations experiencing varying degrees of self-diagnosed mental health disorder symptoms, especially anxiety disorders and major depressive disorder. However, currently published research has yet to explore the possibility of AAT as a supplementary treatment for professionally diagnosed clinical populations. This gap in existing knowledge may be crucial to the future of chatbots in mental health, as the potential findings therein could determine whether or not AAT is a feasible supplement for psychotherapeutic treatment in professional settings. That being said, there is high potential for further research and development on virtual agents in mental healthcare.

Conclusion

The objective of the present study is to serve as a review of artificial intelligence technologies (i.e. social chatbots) as a feasible method of treatment for adult clinical populations suffering from mental disorders. Although a consensus has yet to be reached on the feasibility of autonomous virtual agents in psychotherapy, recent findings show that the concept is indeed a possibility; however, the efficacy of these treatments is currently limited to short-term treatment for certain self-diagnosed mental disorders, namely anxiety disorders and major depressive disorder. As such, current artificial intelligence technologies cannot serve as full replacements for human-administered psychotherapeutic treatment. However, as artificial intelligence technology continues to evolve over short periods of time, the potential for autonomous computerized agents in mental healthcare is a strong possibility worthy of further consideration and analysis.

References

American Psychological Association. (2021). *Worsening mental health crisis pressures psychologist workforce*. American Psychological Association. Retrieved May 22, 2022, from https://www.apa.org/pubs/reports/practitioner/covid-19-2021

Barrett, M. S., Chua, W.-J., Crits-Christoph, P., Gibbons, M. B., & Thompson, D. (2008). Early withdrawal from mental health treatment: Implications for psychotherapy practice. *Psychotherapy: Theory, Research, Practice, Training*, *45*(2), 247–267. https://doi.org/10.1037/0033-3204.45.2.247

Beatty, C., Malik, T., Meheli, S., & Sinha, C. (2022). Evaluating the therapeutic alliance with a free-text CBT conversational agent (WYSA): A mixed-methods study. *Frontiers in Digital Health*, *4*. https://doi.org/10.3389/fdgth.2022.847991

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health*, *4*(2). https://doi.org/10.2196/mental.7785

Foley, G. N., & Gentile, J. P. (2010). Nonverbal Communication in Psychotherapy. *Psychiatry (Edgmont)*, *7*(6), 38–44.

He, L., Basar, E., Wiers, R. W., Antheunis, M. L., & Krahmer, E. (2022). Can Chatbots help to motivate smoking cessation? A study on the effectiveness of motivational interviewing on engagement and therapeutic alliance. *BMC Public Health*, *22*(1). https://doi.org/10.1186/s12889-022-13115-x

Hortensius, R., Hekele, F., & Cross, E. S. (2018). The perception of emotion in artificial agents. *IEEE Transactions on Cognitive and Developmental Systems*, *10*(4), 852–864. https://doi.org/10.31234/osf.io/ufz5w

Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (WYSA) for digital mental well-being: Real-World Data Evaluation Mixed-Methods Study. *JMIR MHealth and UHealth*, *6*(11). https://doi.org/10.2196/12106

Jang, S., Kim, J.-J., Kim, S.-J., Hong, J., Kim, S., & Kim, E. (2021). Mobile app-based chatbot to deliver cognitive behavioral therapy and psychoeducation for adults with attention deficit: A development and feasibility/usability study. *International Journal of Medical Informatics*, *150*, 104440. https://doi.org/10.1016/j.ijmedinf.2021.104440

Leo, A. J., Schuelke, M. J., Hunt, D. M., Miller, J. P., Areán, P. A., & Cheng, A. L. (2022). Digital Mental Health Intervention Plus usual care compared with usual care only and usual care plus in-person psychological counseling for orthopedic patients with symptoms of depression or anxiety: Cohort study. *JMIR Formative Research*, *6*(5). https://doi.org/10.2196/36203

Linardon, J., Fitzsimmons-Craft, E. E., Brennan, L., Barillaro, M., & Wilfley, D. E. (2018). Dropout from interpersonal psychotherapy for Mental Health Disorders: A systematic review and meta-analysis. *Psychotherapy Research*, *29*(7), 870–881. https://doi.org/10.1080/10503307.2018.1497215

Lucas, G. M., Gratch, J., King, A., & Morency, L.-P. (2014). It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, *37*, 94–100. https://doi.org/10.1016/j.chb.2014.04.043

Mishra, P. (2006). Affective Feedback from Computers and its Effect on Perceived Ability and Effect: A test of the Computers as Social Actor Hypothesis. *Journal of Educational Multimedia and Hypermedia*, *15*(1), 107–131.

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Conference Companion on Human Factors in Computing Systems - CHI '94*. https://doi.org/10.1145/259963.260288

Prochaska, J. J., Vogel, E. A., Chieng, A., Kendra, M., Baiocchi, M., Pajarito, S., & Robinson, A. (2021). A therapeutic relational agent for reducing problematic substance use (Woebot): Development and usability study. *Journal of Medical Internet Research*, *23*(3). https://doi.org/10.2196/24850

Ramseyer, F., & Tschacher, W. (2011). Supplemental material for nonverbal synchrony in psychotherapy: Coordinated body movement reflects relationship quality and outcome. *Journal of Consulting and Clinical Psychology*. https://doi.org/10.1037/a0023419.supp

Sharf, J., Primavera, L. H., & Diener, M. J. (2010). Dropout and therapeutic alliance: A meta-analysis of adult individual psychotherapy. *Psychotherapy: Theory, Research, Practice, Training*, *47*(4), 637–645. https://doi.org/10.1037/a0021175

Swift, J. K., & Greenberg, R. P. (2012). Premature discontinuation in adult psychotherapy: A meta-analysis. *Journal of Consulting and Clinical Psychology*, *80*(4), 547–559. https://doi.org/10.1037/a0028226

Weisband, S., & Kiesler, S. (1996). Self disclosure on computer forms. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Common Ground - CHI '96*. https://doi.org/10.1145/238386.238387

World Health Organization. (2018, June 6). *Mental health: Massive scale-up of resources needed if global targets are to be met*. World Health Organization. Retrieved May 22, 2022, from https://www.who.int/news/item/06-06-2018-mental-health-massive-scale-up-of-resources-needed-if-global-targets-are-to-be-met