wysa

Conversational AI for Mental Health: Potential & Risks

The transformative potential of ChatGPT and upcoming conversational AI technologies

Suggested citation. Conversational AI for Mental Health: Potential & Risks: Wysa Inc: 2023

DISCLAIMER

This content is provided for general information purposes only.

This document is produced by the Wysa team. The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of Wysa concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate border lines for which there may not yet be full agreement.

All reasonable precautions have been taken to verify the information contained in this publication. However, the published material is being distributed without warranty of any kind, either expressed or implied. The responsibility for the interpretation and use of the material lies with the reader. In no event shall Wysa be liable for damages arising from its use.

ACKNOWLEDGEMENTS

Leadership: Jo Aggarwal, Ramakant Vempati, Chaitali Sinha

Editorial: Madhavi Roy, Siddharth Thakeria, Meheli Saha, Dyuthi Dinesh, Sarah Baldry

Design: Sam Leonard

Contents

Foreword

Every new technology goes through a cycle of hyperbole, but few have caught people's imagination like ChatGPT. In January this year, the world's leaders gathered in Davos and the most frequent question I heard asked was, "Will ChatGPT take my job?" The hype quickly turned to dismissal as the world realized that the best AI came with its own risks and was prone to "digital hallucinations". Serious interest in using the power of conversational artificial intelligence has only increased since then. And rightly so. While ChatGPT may not be able to directly replace anyone's job just yet, all our roles are likely to change in the next few years as a result of the advances in AI it represents. It is no longer a technology that just reassures us of our own advancement of skills. Nor is it limited to the automation of everyday tasks. It has raised the bar on both the quality and efficiency with which we can operate and has broken notions that humans will always be superior to AI at communicating with other humans.

For us at Wysa, this is not news. While attempting to locate a response to scale and access in mental health, we looked at how technology could meet the need for support that deeply rests on the ability to feel heard and guided without stigma or limits. Since its inception nearly seven years ago, Wysa has held the view that conversational AI as the first step of care offers perhaps the only equitable and systemic solution to the global mental health crisis. Research shows that people are more likely to open up to AI than humans[1] and that AI-guided mental health support and health coaching can create a bond and has efficacy comparable to that of human therapists.

In exploring this use case with AI, we encountered many of the challenges that the world is now grappling with relating to the risks of using AI in sensitive and regulated settings. We developed significant guardrails to leverage AI while achieving 'Best of' mental health apps from privacy champion Mozilla Foundation, as well as recognition from ORCHA[2] for having the highest level of clinical safety.

The learnings from Wysa's journey in putting guardrails around conversational AI to ensure ethical standards, clinical validation, privacy, safety, and transparency are relevant, not only for those seeking easier, better, faster models of delivering mental health support, but also for any organization looking to leverage technologies such as ChatGPT in healthcare and other regulated settings.

In this paper, we discuss what the world has already learnt in using conversational AI for mental health, and the opportunities and risks around using new technologies such as ChatGPT. We hope to leave you with a frame within which to review these risks and decide what guardrails your organization may need in order to leverage ChatGPT in a regulated setting.

Jo Aggarwal
Founder and CEO
Wysa

Where we are today

Given the current global crisis in mental health, the integration of AI in the mental health ecosystem is leading to transformative developments across the world. AI has made it easier for mental health professionals[3] to refine and confirm their diagnoses, recognize early patient symptoms, and customize treatments based on individual characteristics. Virtual health assistants and chatbots are promoting patient autonomy and self-management[4].

AI has made it possible to identify mental health concerns from text patterns, as well as verbal and non-verbal indicators of psychological distress[5]. It can also coalesce information from multiple data streams to predict health risk[6], can identify trends in longitudinal data, and improve diagnostic precision. AI is also assisting clinicians in monitoring symptoms[7] and administering[8] psychometric assessments.

From the point of view of users and patients, the agility of AI-guided mental health support has allowed it to be usable, meaningful, and relevant to needs across diverse populations.

The solution's ability to be accessed on a virtual or mobile platform helps address an essential part of inclusivity and equity in access. For instance, individuals living in rural communities or shift workers may not have access to mental health facilities at times or locations that suit their needs, and an AI chatbot can be a potential solution. An equally valid use case is those who need support intermittently throughout the day to get through a difficult period.

AI-based mental health support as the first step of care is arguably the only scalable systemic solution to the global mental health crisis. With over half the world living in areas with less than one psychiatrist for every 250,000 people, and the presence of long waiting lists and resource constraints even in developed economies, solutions like Wysa have used conversational AI to deliver therapeutic support that bridges key gaps in healthcare provision.

In the current state of play, conversational AI for mental health is being embraced by users who appreciate having something to talk to and guide them when there is no other support available. It is being welcomed by large employers who understand the benefits of a mentally well workforce, and by healthcare providers to support patients across different stages of care.

The Wysa experience shows the power of conversational AI when the right guardrails are in place. We have seen AI deliver significant clinical outcomes while setting standards of privacy and safety. We have also seen surprising results in user acceptance of conversational AI as a coach; not merely as a low-cost version of human support but as a completely different entity with its own unique advantages. In the next two sections, we delve deeper into these.

The Wysa Experience: Clinical Efficacy, Safety, and Privacy

With all the hype about ChatGPT, it is easy to think of the next generation of conversational AI as a party trick that could write a term paper. The real impact of AI, however, is when it is invisibly working in the background of a solution that delivers clinical outcomes, acting as an amplifier that allows such solutions to scale up without the need for human moderation, and deliver access to care at a cost not possible without it.

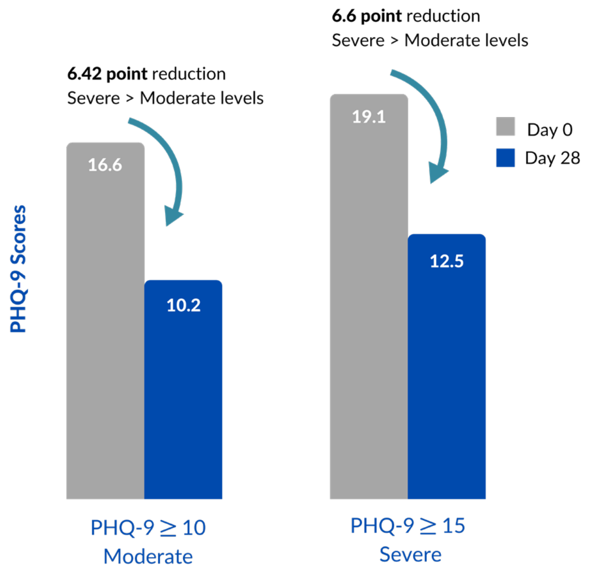
This is how Wysa has used conversational AI; as a layer that allows people to talk in natural language about their issues to feel heard and supported, alongside clinically approved algorithms that deliver coaching and support.

Used in this way, Wysa's AI has been shown to reduce symptoms of depression and anxiety and develop a therapeutic alliance on par with human therapists in a shorter period of time[9]. Wysa has delivered positive clinical outcomes across age groups and across both mental and physical functions. Most recently it was shown to reduce psychological distress and pain interference[10,11] in older adults with chronic pain, for which it has received the Food and Drug Administration (FDA) breakthrough device designation[12].

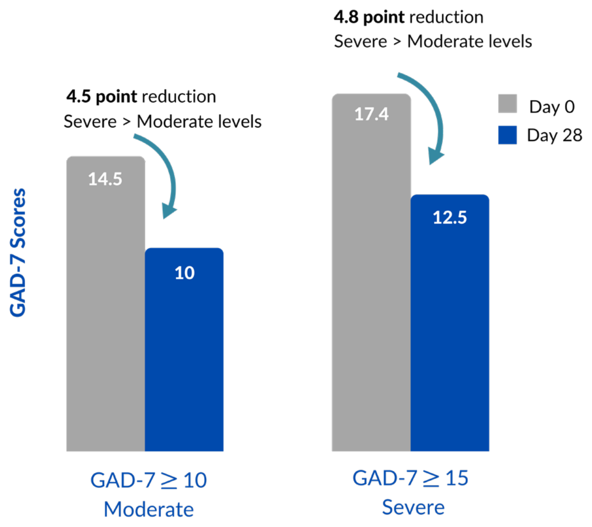
A key part of Wysa's appeal stems from its ability to use natural language AI and create an empathetic interface that is highly engaging and effective for users while retaining guardrails that make the platform safe, private, and clinically validated.

From a HIPAA perspective, Wysa minimizes any data collection and redacts users' accidental sharing of personal identifiable information (PII) to ensure data protection. It complies with NHS DCB 0129 Risk Management Standards, and with ISO 27001 & 27701 to ensure the standardization of privacy and security standards across the product. Wysa's privacy prioritization and standards led to privacy watchdog Mozilla Foundation praising Wysa as the 'Best of' mental health apps and one of the highest ratings on clinical safety (93%) based on data security, clinical assurance, and user experience by the Organization for the Review of Care and Health Apps (ORCHA)[13].

Wysa's efficacy with an insurer supporting 60,000 members in their well-being journey



Improvements in symptoms of depression after 28 days



Improvements in symptoms of anxiety after 28 days

The Wysa Experience: Usability, Acceptability, and Engagement

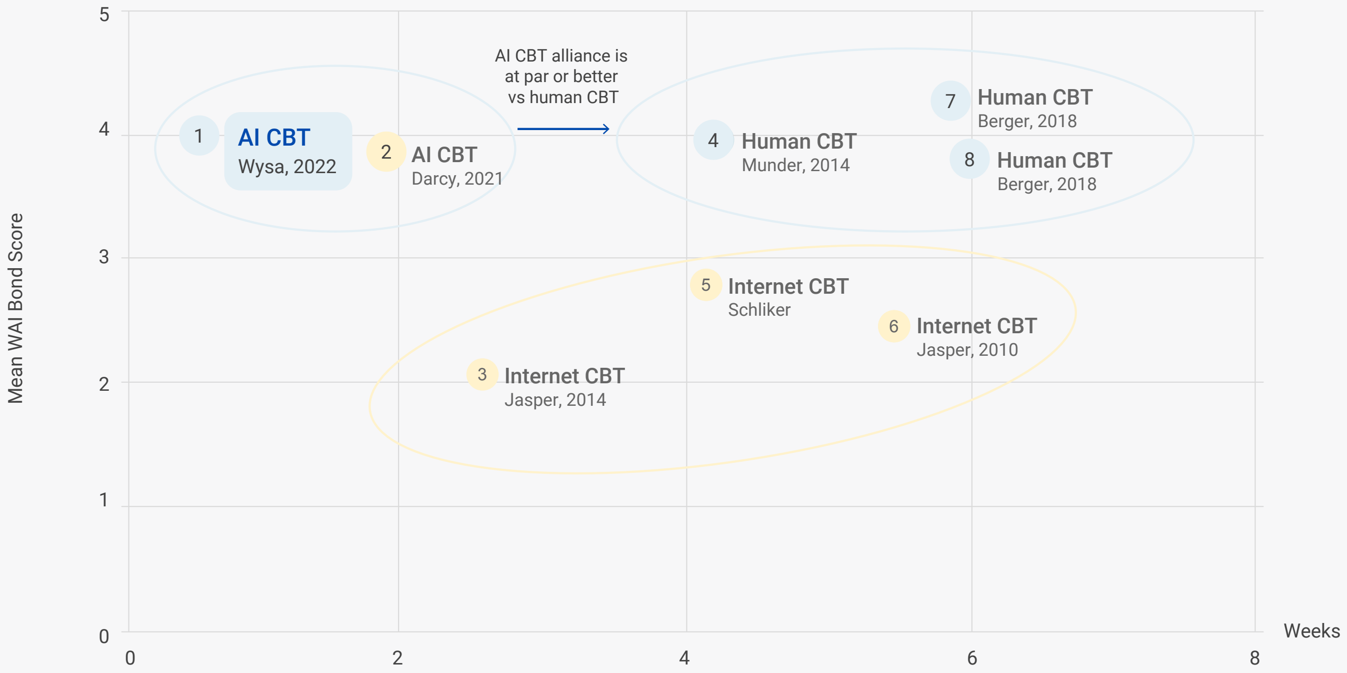
Is it possible to create a combination of clinical logic, empathetic design, and AI for conversation and self-disclosure where the user can feel heard and be guided through evidence-based techniques with AI alone, with no human in the conversation? This is the question that we set out to answer in the early days of Wysa. Now, having had over half a billion conversations with six million users across the world, we can say with confidence that conversations with AI are seen as safe spaces with their own unique characteristics. These characteristics, in fact, make them even more suitable for coaching, just-in-time support, and disclosure, than human support.

To understand how AI can engage users, we conducted a study of Wysa's user acceptance on indicators of usability, usefulness, and acceptability, as defined by the Healthcare Information and Management Systems Society (HIMSS) framework, with feedback from well over 40,000 individuals. Wysa's AI platform rated highly, with users indicating that the AI conversations felt safe, engaging and non-judgemental[14].

People may enjoy talking about their mental health worries to AI, but can it guide them to change behaviors? In another study, we explored this question using the Working Alliance Inventory (WAI-SR). This measures the therapeutic bond between the AI and the user and is an indicator of how helpful AI is perceived to be and how effective the AI would be as therapeutic support. We found that individuals who used Wysa developed a robust therapeutic alliance with the AI conversational agent as soon as 3-5 days, at a level comparable to an alliance with a human therapist[15]. Qualitative analysis has further documented the positive impact of dynamic free-text within Wysa that mirrors the capacity within human interactions for reciprocity, engagement, and acknowledgment.

Establishing a Therapeutic Alliance with AI-CBT

Therapeutic Alliance of AI-CBT is on par with Human CBT and is established within 3-5 days



These findings show the power of conversational AI to deliver clinical outcomes and efficiencies, but to many, they may still appear conceptual. To truly appreciate how Wysa's conversational AI is changing lives in the real world, let us look at Chukurah's story, as told to NPR[16].

The Real-World Impact of Conversational AI

NPR recently reported the lived experience of a woman named Chukurah Ali, who significantly benefited from Wysa, evidencing the human experience of AI's ability to bridge the key gaps in existing care. Chukurah, a single mom and employer running a successful bakery business, Coco's Desserts in St Louis, lost her business after a car accident left her injured and unable to walk.

> "
>
> *I could barely talk. I could barely move. I felt like I was worthless because I could barely provide for my family at that moment. And now I lost my car. I can't even take care of my daughter.*

She sank into depression, feeling hopeless and worthless. When it came to finding help, she couldn't find an available therapist, let alone drive her car to get to appointments. She was also left bereft of her insurance. Her orthopedic doctor at Washington University recommended Wysa, propelling Chukurah's recovery journey. During the first few months, she talked to the AI bot almost every day, sometimes as much as seven times a day.

> "
>
> *I thought it was weird at first," she said. "I'm like, OK, I'm talking to a bot. It's not going to do nothing. I want to talk to a therapist. But that bot helped! I would just start chatting with it. 'How are you feeling today?' Then it would give me these little options that I could do.*
>
> *What I noticed it was doing - CBT therapy, cognitive behavioral therapy. It's not a person but it makes you feel like it's a person because it's asking you all the right questions.*

In 8 weeks, there were significant shifts in both pain and mood. One year on, she continues using Wysa alongside human healthcare support.

Coco Ali
Coco's Desserts

How does ChatGPT change things?

ChatGPT has begun to fundamentally change how people perceive human-computer interactions, with implications for both the provider and the patient.

Just as Google Search fundamentally changed[17] the nature of doctor-patient conversations with the widespread availability of information, the evolution of conversational AI will change that interaction further. The ability to synthesize data across medical reports, the creation of continuous medical support for chronic conditions, and the improvement of systemic responsiveness can elevate the patient experience. Secondly, clinical decision-making and error minimization can improve further with AI support.

Given these expectations, and that the presence of this technology will start becoming more normative, the ability to leverage ChatGPT reliably and safely in a highly regulated setting will become key.

While users will expect healthcare providers to leverage these technologies for a better patient experience, regulators and clinicians are right to be cautious.

Artificial Intelligence is prone to bias and errors of judgment just like humans, but the organizational liability of AI errors can be significantly higher. When replacing a human-moderated environment with an AI-moderated one, there is little tolerance for issues arising from AI's risks. The risks themselves are numerous.

There have been reports of AI chatbots being hostile to users, passing "snide remarks and angry retorts,"[18] and AI systems such as GPT, GPT-2, and GPT-3 have generated at least 1 in 100 responses containing toxic and abusive language[19].

Especially in a regulated setting, AI-powered generative text responses can be potentially harmful, as they are not clinically vetted or explainable.

Perhaps the first barrier to the adoption of ChatGPT by healthcare organizations is their vendor engagement forms, which start with ensuring compliance by stipulating privacy regulations such as HIPAA and GDPR.

Addressing Risks with Guardrails

There is tremendous potential in ChatGPT. Patients and clinicians will expect to leverage this technology, yet the risks and regulatory constraints are significant. How can one frame this challenge in a way that allows us to move forward and harness this potential while mitigating risks?

Our experience at Wysa helps us create a frame within which to view the many alarming news stories that you may encounter in the hype cycle of ChatGPT and similar technologies, while still having a clear answer on how healthcare organizations can manage these risks.

Instead of individual risks, at Wysa we look at the components of AI that generate such risks. For ChatGPT, there are three main sources of risks in using it in a regulated setting.

1. **The generative nature of AI**
2. **Sharing data with third parties**
3. **The accuracy and reliability of AI**

Let's look at each of these risks and how Wysa has successfully mitigated them.

1

Firstly, there are risks emanating from the generative nature of its AI. Any generative AI will produce different text each time, making it difficult to test for clinical safety or validity. To address this risk, one needs to limit how generative AI is used within the organization.

At Wysa, we limited the use of AI to natural language understanding and personalization within a clinically approved algorithmic framework. Generative capabilities, where used, are adjunctive to this framework and are explainable, testable, and clinically approved.

Wysa's framework to deliver responses to users is a rule-based system created by clinicians that leverages AI to converse using natural language. For instance, AI can help determine if a user's statement contains suicidal ideation, but the response to that statement, which includes an escalation path to helplines and the technique for working on a safety plan, is driven by an explainable, auditable, and clinically approved algorithm. The usage of AI in this manner can be validated and tested for safety, something that would not be possible with a generative AI response, no matter how appropriate it was.

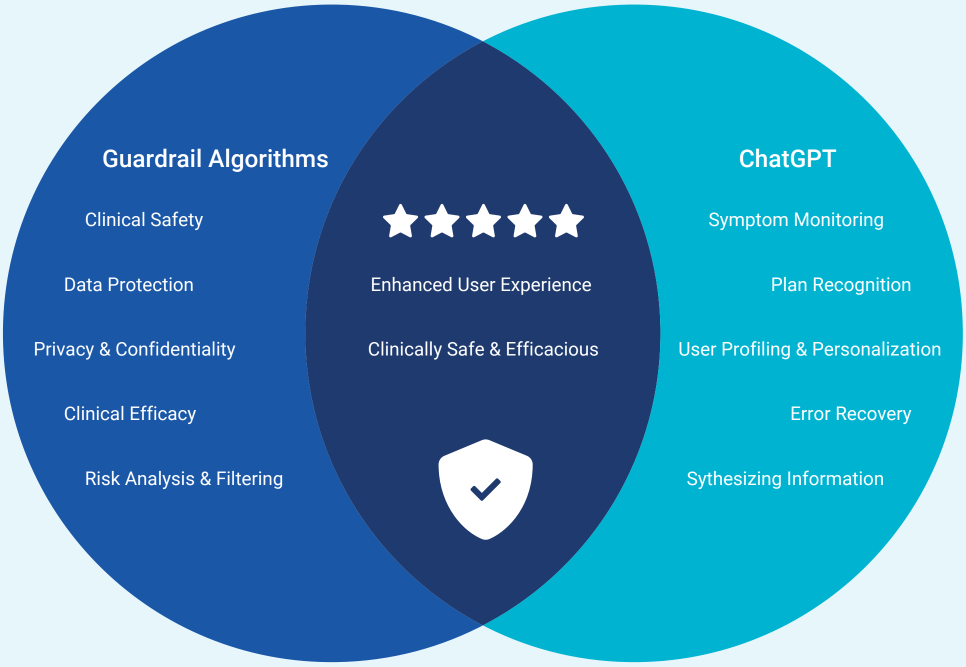
A rule-based framework built by clinicians creates the guardrails to introduce some elements of generative AI to provide a human-like conversational experience, for instance, to improve user engagement and gather more context, while being evidence-based in output.

How does ChatGPT change things?

2 The second key risk stems from sending data out to a third party: this leads to privacy, security, and regulatory risks. A first step for any regulated organization seeking to use ChatGPT would be to replace all PII with placeholders or synthetic PII that will allow ChatGPT to provide accurate results without storing any personally identifiable information. After receiving the response from ChatGPT, there would then be a layer at the organization's end that replaces the synthetic PII in the output with the original PII so that the output is meaningful to a patient or clinician.

3 The third category of risk stems from the accuracy and reliability of artificial intelligence. Even the best AI will fail in ways that are hard to explain. A key element of clinical safety is to be able to test a product and ensure that even when the AI failure leads to a less-than-appropriate response, such a response does not have the potential to cause harm to a patient or user.

Managing this risk requires rigor, discipline, and humility. At Wysa we test each piece of content for scenarios where AI fails to detect appropriate risk and look at the response the user would receive when the AI fails. This is then manually assessed, and failsafe responses are designed, so that if the AI fails, the response is still not triggering or encouraging harmful behavior. This is possible because Wysa's content is not generative - only its natural language understanding and profiling uses AI.



Implementing guardrail algorithms on technologies like ChatGPT will elevate the clinician-and-user experience in the digital mental health space.

Each of these guardrail algorithms and restraints will be needed when organizations adopt ChatGPT for patient interactions. They may appear to 'dumb down' the capabilities of AI, but we can see from Wysa's experience that the priority is to be able to remove access barriers while ensuring that clinical safety and efficacy can have a transformative impact on some of our biggest healthcare challenges.

Conclusion

Looking beyond the hype of ChatGPT, conversational AI technologies can help us solve some of the world's most pressing health crises.

With Wysa, we have already seen its potential in addressing mental health and chronic conditions. Adding conversational AI as the first step in the care continuum can help bridge the shortage of qualified professionals, and remove barriers to access related to stigma, cost, and availability. AI can build a therapeutic alliance and create equitable access to support at scale.

The breakthroughs in artificial intelligence in general, and ChatGPT in particular, are inspiring us to dream of a new world, and with it, changing cultural discourse. The hype of these technologies has been quickly followed by skepticism as their vulnerabilities and risks are revealed.

In this report, we used Wysa's journey in using conversational AI as an enabler for solving the global mental health crisis as a guide on how healthcare organizations might navigate the opportunity that ChatGPT presents. We looked at ways in which conversational AI can be used in regulated settings, and areas where it may need guardrails and constraints. Most importantly, through Wysa's own experience, we hope to show ways in which AI can overcome regulatory challenges, create user acceptance and engagement, and deliver clinical efficacy at scale to solve the world's biggest health challenges.

Wysa's roadmap and guardrails have enabled it to remove barriers to access to mental health and deliver clinical outcomes at scale. We hope these learnings will also inspire other organizations to leverage technologies like ChatGPT in ways that are safe, private, and transformative for patients and clinicians.

References

1. Berger BA. Building an effective therapeutic alliance: Competence, trustworthiness, and caring. Am J Health Syst Pharm. 1993 Nov 1;50(11):2399–403.

2. Mental health apps rating map [Internet]. ORCHA. 2019 [cited 2023 Feb 23]. Available from: https://orchahealth.com/mental-health-apps-rating-map/

3. Graham S, Depp C, Lee EE, Nebeker C, Tu X, Kim HC, et al. Artificial Intelligence for Mental Health and Mental Illnesses: an Overview. Curr Psychiatry Rep. 2019 Nov 7;21(11):116.

4. Callejas Z, Griol D. Conversational Agents for Mental Health and Wellbeing. In: Lopez-Soto T, editor. Dialog Systems: A Perspective from Language, Logic and Computation. Cham: Springer International Publishing; 2021. p. 219–44.

5. Rizzo A, Shilling R, Forbell E, Scherer S, Gratch J, Morency LP. Chapter 3 - Autonomous Virtual Human Agents for Healthcare Information Support and Clinical Interviewing. In: Luxton DD, editor. Artificial Intelligence in Behavioral and Mental Health Care. San Diego: Academic Press; 2016. p. 53–79.

6. Marmar CR, Brown AD, Qian M, Laska E, Siegel C, Li M, et al. Speech-based markers for posttraumatic stress disorder in US veterans. Depress Anxiety. 2019 Jul;36(7):607–16.

7. di Biase L, Raiano L, Caminiti ML, Pecoraro PM, Lazzaro VD. Artificial intelligence in Parkinson's disease—symptoms identification and monitoring. In: Pillai AS, Menon B, editors. Augmenting Neurological Disorder Prediction and Rehabilitation Using Artificial Intelligence. San Diego, CA: Elsevier; 2022. p. 35–52.

8. D'Alfonso S. AI in mental health. Curr Opin Psychol. 2020 Dec;36:112–7.

9. Inkster B, Sarda S, Subramanian V. An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. JMIR Mhealth Uhealth. 2018 Nov 23;6(11):e12106.

10. Leo AJ, Schuelke MJ, Hunt DM, Metzler JP, Miller JP, Areán PA, et al. A Digital Mental Health Intervention in an Orthopedic Setting for Patients With Symptoms of Depression and/or Anxiety: Feasibility Prospective Cohort Study. JMIR Form Res. 2022 Feb 21;6(2):e34889.

11. Meheli S, Sinha C, Kadaba M. Understanding People With Chronic Pain Who Use a Cognitive Behavioral Therapy-Based Artificial Intelligence Mental Health App (Wysa): Mixed Methods Retrospective Observational Study. JMIR Hum Factors. 2022 Apr 27;9(2):e35671.

12. Wysa Receives FDA Breakthrough Device Designation for AI-led Mental Health Conversational Agent [Internet]. 2022 [cited 2023 Feb 23]. Available from: https://www.businesswire.com/news/home/20220512005084/en/Wysa-Receives-FDA-Breakthrough-Device-Designation-for-AI-led-Mental-Health-Conversational-Agent

13. Mental health apps rating map [Internet]. ORCHA. 2019 [cited 2023 Feb 23]. Available from: https://orchahealth.com/mental-health-apps-rating-map/

14. Malik T, Ambrose AJ, Sinha C. Evaluating User Feedback for an Artificial Intelligence-Enabled, Cognitive Behavioral Therapy-Based Mental Health App (Wysa): Qualitative Thematic Analysis. JMIR Hum Factors. 2022 Apr 12;9(2):e35668.

15. Beatty C, Malik T, Meheli S, Sinha C. Evaluating the Therapeutic Alliance With a Free-Text CBT Conversational Agent (Wysa): A Mixed-Methods Study. Front Digit Health. 2022 Apr 11;4:847991.

16. Noguchi Y. Therapy by chatbot? The promise and challenges in using AI for mental health. NPR [Internet]. 2023 Jan 19 [cited 2023 Mar 14]; Available from: https://www.npr.org/sections/health-shots/2023/01/19/1147081115/therapy-by-chatbot-the-promise-and-challenges-in-using-ai-for-mental-health

17. Van Riel N, Auwerx K, Debbaut P, Van Hees S, Schoenmakers B. The effect of Dr Google on doctor-patient encounters in primary care: a quantitative, observational, cross-sectional study. BJGP Open. 2017 May 17;1(2):bjgpopen17X100833.

18. Ray S. Bing Chatbot's "Unhinged" Responses Going Viral. Forbes Magazine [Internet]. 2023 Feb 16 [cited 2023 Feb 21]; Available from: https://www.forbes.com/sites/siladityaray/2023/02/16/bing-chatbots-unhinged-responses-going-viral/

19. Gehman S, Gururangan S, Sap M, Choi Y, Smith NA. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models [Internet]. arXiv [cs.CL]. 2020. Available from: http://arxiv.org/abs/2009.11462

wysa

www.wysa.com