# Key Considerations of Artificial Intelligence in Cognitive Behavioural Therapy

Marcus Chad

## ABSTRACT

The integration of artificial intelligence (AI), particularly conversational AI (CAI), into cognitive behavioural therapy (CBT), is increasingly proposed as a scalable response to the global mental health treatment gap. However, this technological promise introduces a fundamental tension with the core pillars of CBT, which rely on a genuine therapeutic alliance and ethical accountability. This paper aims to examine the efficacy and safety of AI-CBT across dimensions of client satisfaction, clinical safety and moral agency. The analysis reveals a double-edged model: features that boost AI's scalability and promise in the field also reveal underlying tensions with building a therapeutic alliance and ethical accountability. The perceived anonymity, whilst boosting initial patient disclosure and engagement, does not allow for genuine rapport to be built. The sycophantic behaviours of AI, despite again boosting initial engagement rates, may undermine the development of a therapeutic alliance. Furthermore, unregulated AI poses severe clinical risks, evidenced by documented ethical violations and the emerging reports of AI-psychosis, a proposed phenomenon where AI usage leads to psychosis. Consequently, this paper argues that AI tools must be subjected to rigorous regulatory assessments rather than rapid consumer deployment. The evidence supports a hybrid clinician-AI model as the most defensible framework, where data intensive tasks are delegated to the AI to allow the therapist to focus on building a therapeutic connection.

## INTRODUCTION

The integration of artificial intelligence (AI) into cognitive-behavioural therapy (CBT) is often framed as a potential solution to the global mental health treatment gap. AI promises scalable, low cost, always-available support. In psychological treatment, conversational AI (CAI) based on Large Language Models (LLMs) are increasingly being considered as an alternative to talk-therapy. They operate by converting the client's input into tokens, which the model processes to generate the statistically most probable sequence of words, given the distribution of its training corpus (Shanahan et al., 2023). Yet this promise stands in tension with the foundations of CBT: a collaborative process whose efficacy depends on both a genuine therapeutic alliance and ethical accountability.

CBT is a form of psychological treatment that helps individuals identify and develop strategies to challenge negative thought patterns and unhelpful behaviours, aiding them to identify the intersections between their actions and feelings (Corliss et al., 2024). It is an evidence-based therapy that has been demonstrated to be effective for a range of problems including depression, anxiety disorders, and severe mental illnesses. CBT is widely considered a treatment of choice for several mental health conditions in the psychological space (Corliss et al., 2024).

A therapeutic alliance is defined as the collaborative and trusting relationship between a therapist and a client characterised by mutual agreement to collaborate on tasks related to the patient's well-being, which is considered a crucial factor for positive treatment outcomes (Nadja et al., 2022). Clients have demonstrated satisfaction with AI, as qualitative studies suggest

(Hatch et al., 2025). However, there are several factors that prevent the development of a therapeutic alliance dependent on genuine rapport. AI typically demonstrates sycophantic behaviours, which means that the AI is overly agreeing and affirmative of the patient, which builds a false facade of intimacy (Ifthikar et al., 2025). Furthermore, patients do not feel the same way they do with AI as they do with real people, as AI lacks moral agency. In the case of the *Paro* robot, an early application of AI into the field of psychology to treat dementia, certain patients noted that they were clearly aware that it was a fraud (Griffiths et al., 2014). The feeling of the involvement of AI in CBT could potentially make patients feel the same way; as if it was a fraud. However, the implementation of *Paro* into this study of elderly people was not entirely negative. The *Paro* robot's ability to comfort patients demonstrates that the use of AI into psychological care is highly nuanced (Griffiths et al., 2014).

As concerns around AI-CBT continue to arise, researchers have attempted to create concrete guidelines to follow when using it in clinical psychology, reflected by the rise of AI ethics counselors and discussions of teaching ethics to data scientists (Ewbank et al., 2020). Notably, debate continues over whether AI, despite its shortcomings, should be included into CBT to increase the effectiveness of treatment. This paper argues that AI should undergo rigorous assessment before being integrated into CBT, and that integration should proceed only once its potential benefits and drawbacks are more clearly understood.

## CLIENT EXPERIENCE

As AI moves from a peripheral support role to a central cornerstone for delivering CBT, evaluating AI's potential in CBT is crucial, and a comprehensive comparison must be made. In this section, we provide an assessment of AI-CBT across the main dimensions of: 1. Client satisfaction, where the quality of the therapeutic relationship that develops is examined; and 2. Client usability, where the engagement, and acceptability of AI-CBT tools are assessed. Through the assessment of these sections, the double-edged nature of AI within this field will become evident. Whilst features may provide benefits such as boosting engagement rates or building rapport, those same features may prove disadvantageous upon closer inspection.

Firstly, we must consider the client satisfaction of AI-CBT. In one study, researchers examined clients who used *Clare*, a voice-enabled AI chatbot designed for mental health support. (Note, the lead researcher of this study was dually employed by Clare.) The study examined what clients expected from *Clare*, their motives for usage, and effects on clients after initial usage of the chatbot.(Schafer et al. 2025) The study found that users were attracted and bonded with AI-CBT for self-help, seeking "emotional support without concerns about shame or physical appearance." (Schafer et al. 2025) Despite fears regarding a lack of trustworthiness and intimacy from going through an online process with an AI chatbot arising, especially because it is merely reproducing statistically probable responses; clients actually found this approach in using AI-CBT helpful because it removed any sense of judgement.(Schafer et al. 2025) Despite the dual employment of the authors of the *Clare* study, some of their results do collude with independent studies. Particularly, a result both studies showed was that participants were willing to disclose information to a chatbot as much as a human (Croes et al. 2024). As people feel more anonymous, their public self-awareness decreases, which reduces identifiability or accountability concerns. This can result in more intimate self-disclosure. A sense of anonymity is easily established with CAI since the client is discussing with an artificial being rather than a human, removing a sense of judgement. Croes et al. notes that the fear of

judgement may hinder humans disclosing information to other humans, but this is not applicable to CAI due to its perceived anonymity, as CAI is unable to form judgements on its own (Croes et al., 2024). Overall, the study by Croes et al. found that perceived anonymity was the only variable that directly predicted the disclosure. A separate study by Cheng et al. (2024), conducted across a three-day music festival, adds a key nuance. Despite the perceived anonymity when disclosing information to CAI, there was less trust in the CAI in comparison to when disclosing information to a human. Furthermore, the self-disclosure intimacy was higher for humans than CAI (Cheng et al., 2024). This lack of trust in CAI may have arisen due to its perceived anonymity, reflecting a double-edged sword effect. Therefore, despite hopeful results arising from Cheng et al.'s and Schafer et al.'s study in relation to an increased likelihood of disclosure of information by patients, this does not reflect a building of genuine rapport and therapeutic alliance as there is still a lack of trust, required in CBT for effective treatment. Furthermore, the snapshot nature of Cheng et al.'s study raises concerns. It happened across a three day music festival, where the relationships of participants with the AI companion and other humans are more surface-level (Cheng et al., 2024). Disclosing to a therapist occurs within a structured professional relationship, and Cheng et al. does not replicate that within its study (Corliss et al., 2024).

Secondly, despite chatbots not having moral agency potentially being a benefit, it can be double-edged as they can spread misinformation and be sycophantic. They are only able to repeat what statistics tell them, making them unreliable.(Croes et al. 2024) AI typically exhibits overly sycophantic behaviour, reducing its ability to effectively provide proper, effective treatment. Sycophantic behaviour from AI has been shown to lead to higher engagement rates, leading to a perverse incentive for this behaviour to persist in AI (Cheng et al., 2026). Despite distorting judgement, sycophantic models are trusted and preferred. The users who engaged with sycophantic models in the study conducted by Cheng et al. displayed a lack of proper judgement, specifically reducing their willingness to bear responsibility (Cheng et al., 2026). This reflects a potentially destructive path for this behavioural trait of AI, incentivized to persist as it increases user engagement.

Finally, the therapeutic alliance between client and therapist could potentially be broken down by AI. In a study (Xu et al., 2025) evaluating depression intervention with AI chatbots with social cues, it concluded that chatbots with high-social cue designs significantly outperformed text-only versions in alleviating depression and anxiety whilst enhancing adherence, satisfaction and therapeutic alliance. Therapeutic alliance meaning the connection between patient and therapist characterised by mutual agreement to collaborate on tasks related to the patient's well-being (Encyclopedia of Psychotherapy, 2002). These findings suggest that an AI's ability to incorporate social cues is especially important, and results in greater client satisfaction leading to better client outcomes. The 'High Social Cue' design resembles that of an actual person more, and this effect was achieved through incorporating texts with voice notes and animations, which resulted in a greater efficacy of treatment. However, a study from Ifthikar et al demonstrates that AI is not able to consistently match that and pick up on social cues, especially when it demonstrates sycophantic behaviours. AI models, when conversing with patients, generally exhibited poor therapeutic collaboration, where the LLMs generated overly lengthy responses in comparison to the patient. This shifted the tone of the session from a casual environment where the patient felt free to express themselves, to one of a lecture. Additionally, this behaviour also over-validated patients' harmful thoughts about themselves (Ifthikar et al. 2025). To conclude, despite there being methods of AI being able to overcome the barrier of a

lack of intimacy with a patient, due to it being a non-sentient, non-physical being, these findings suggest that, at present, AI-CBT struggles to consistently foster a collaborative and responsive alliance that is central to effective therapy.

Evidence suggests that AI-CBT is broadly usable and initially engaging, but sustained client-friendliness remains fragile (Shanahan et al. 2023). Users consistently rate AI-CBT tools as easy to use, as the process only requires a prompt from the user to function. Users also report moderate to high satisfaction, as the constant access, non-judgemental interaction and practical coping strategies the AI provides are helpful (Croes et al. 2024). However, persistent frustrations with conversational rigidity, lack of genuine personalisation and occasional misunderstandings limited its ability to build genuine rapport and acceptability (Schafer et al. 2025). These may have arisen due to the fact that AI typically accounts for messages independently from each other, preventing their ability to construct an overall bigger picture of the patient and have genuine conversational fluidity (Cheng et al. 2026). Qualitative research maps the key barriers and fears: privacy fears, perceived lack of human contact, AI dependency concerns, and uncertainty about handling complex problems all limit sustained use. When using ChatGPT as an academic support tool, participants of the study noted that there were fears about a loss of competency in foundational skills, risks to academic integrity and raised concerns about AI's reliability and accuracy (Lam et al., 2025). Whilst this does not explicitly link to usage of AI in CBT, these concerns apply to AI generally. The calling for robust institutional policies by participants in Lam et al.'s study is also applicable to AI-CBT. Overall, AI-CBT can briefly boost engagement, but its friendliness depends on design features that overcome the relational warmth users need (rosenstrom et al. 2025).

In conclusion, the integration of AI into CBT presents a complex landscape. The proposed benefits of accessibility, consistency, and stigma reduction are seemingly enough to justify its inclusion. However, upon closer examination, these same benefits that are provided by the features of AI turn to also work against it; AI suffers significant challenges in fostering a genuine therapeutic alliance and sustained client satisfaction. AI-CBT presents a promising response to the mental health treatment gap, but its shortcomings reveal a tension in its integration. The central question is therefore not only whether AI belongs in CBT, but how quickly it should be integrated: whether the urgency of the treatment gap justifies immediate adoption, or whether integration should wait for the comprehensive understanding and rigorous assessment its shortcomings demand.

## CLINICAL RISKS AND REGULATORY OVERSIGHT

The rapid proliferation of artificial intelligence tools, particularly conversational AI and large language models (LLMs), presents a compelling, yet complex, opportunity to address the persistent global shortage of human clinicians and democratize access to mental healthcare. However, this alluring prospect introduces a fundamental tension: the speed of technological integration versus the imperative for rigorous safety assessment. This section critically examines the documented and potential safety failures of unregulated AI in mental health, explores emerging regulatory and ethical frameworks, discussion of its benefits, and conceptualizes a balanced path forward that avoids a false dichotomy between total inaction and reckless speed.

Policymakers are beginning to address these dilemmas through legislative and regulatory action. As of late 2025, the US Food and Drug Administration (FDA) is proactively adopting a

risk-based framework for regulating generative AI in mental health. During a meeting of the Digital Health Advisory Committee (DHAC), the FDA articulated its intent to apply a Total Product Life Cycle (TPLC) approach to generative AI-enabled mental health medical devices (Venable).

The TPLC framework is the FDA's long-standing framework for medical device regulation. The application of this established device lifecycle framework to AI signals that these chatbots in relation to CBT are being treated as medical devices, which are subject to premarket evidence and postmarket monitoring (Venable LLP, 2025) . This approach acknowledges the dynamic nature of AI, where models change over time, necessitating continuous oversight beyond initial market approval. Key considerations within this framework include the importance of physician/human oversight and intervention, recognizing that AI tools should augment, not replace, human clinical judgment.

The empirical justification for such a cautious, evidence-intensive approach is building. A recent study conducted by Iftikhar et al. (2025) systematically evaluated the performance of CBT-prompted LLMs across 110 self-counselling sessions and 27 simulated therapy sessions with an LLM counsellor using publicly available transcripts of human therapy. The findings revealed a consistent pattern of serious ethical violations by the LLMs.

Firstly, the LLMs exhibited rigid methodological adherence, where the LLMs lacked the clinical interpretation to tailor a therapeutic intervention to match a user's context resulting in a blanket intervention. Additionally, the LLMs typically dismiss lived experiences by the user, where it offers oversimplified, generic and context-insensitive advice to users, particularly to those from minority identities (Ifthikar et al. 2025). Secondly, the study documented poor therapeutic collaboration, where the LLMs exhibited poor turn-taking behaviour by generating overly lengthy responses that detract from the user's voice and focusing it on its own, shifting the tone of the session from therapy to a lecture (Ifthikar et al. 2025). Adding onto this, it validated unhealthy beliefs, where the LLMs reinforced users' inaccurate and harmful beliefs about themselves and others, as LLMs do show a pattern of sycophancy as proven by Sharma et al. (2023). Finally, the LLMs showed signs of gaslighting, which is a form of psychological manipulation that causes the victim to question the validity of their own thoughts, as it makes improper correlations between the users' thoughts and behaviours, in some cases incorrectly suggesting that users are causing their own mental health struggles. This behaviour is also intriguing because it is almost paradoxical to the sycophantic behaviour the AI shows, reflecting once more the double-edged nature of AI-CBT, developed upon in section 1. Deceptive empathy was also noted, where the LLMs' 'pseudo-therapeutic' alliance, where the LLM posed as a social companion, included simulated anthropomorphic responses, which may be due to its again, sycophancy, that created a false sense of emotional connection that could've been misleading, especially for vulnerable groups (Ifthikar et al. 2025). Iftikhar et al. also showed that unfair discrimination was prevalent in LLMs' responses; in one example, it flagged discussions involving female perpetrators as violations of terms of service, while similar male-related content does not result in violations. Another example of unfair discrimination that LLMs demonstrated in this report was religious bias, where LLMs mislabels values and practices from minority religions as content endorsing extremism. Overall, the LLMs tended to show bias towards Western practices. Finally, the LLMs demonstrated a lack of safety and crisis management, as the LLMs tended to show poor crisis navigation, as it responded either indifferently, disengages or fails to provide appropriate intervention in crisis, for example when the patient showed suicidal tendencies or self-harm), and the LLMs also demonstrated abandonment in some cases, where

service was completely denied and LLMs sometimes stopped responding to sensitive topics, such as depression (Ifthikar et al. 2025).

Additionally, a new special report produced by Adrian Preda documents the rise of a new proposed phenomenon based on previous case reports: AI-induced psychosis (AIP). In the report, it documents multiple case studies of this phenomenon. One case study being a Belgian man who was diagnosed with schizophrenia and bipolar disorder developing an emotional attachment to ChatGPT and commited suicide after, his name being Taylor (2025). Another case study being that of a Wisconsin man, his name being Jargon (2025) on the autism spectrum who rapidly spiraled into mania after chatbot validation, and Brandon, who after going through a painful breakup develops a strong emotional attachment to a chatbot he names 'Paul', which affirms his paranoid beliefs ultimately making Brandon quit his job and withdraw from society. Across all these cases, mental health risk factors, including loneliness, long hours of uninterrupted chat that led to a poor sleep schedule, and chatbot memory features designed for personalisation, ended up reinforcing delusional themes that ultimately led to mania or psychosis. A review of chat logs by Sharma, et al. (2023) revealed no attempts by the chatbots to refute the users' delusions or perform risk assessments. The report links 'monomania', a fixation on a single mental subject, to AIP where the idea, or mental subject that the patient is fixated on is an AI companion. The report documents that symptoms of AIP include psychotic symptoms that overlap with mood changes, poor judgement, limited insights, neurovegetative symptoms, and behavioural changes, along with delusions centered on the theme of the AI companion being sentient, possibly being present. The report identifies 5 potential risk factors that the reported cases appear to share. Firstly, intensive use of the chatbot can lead to the development of AIP. Most rare and severe cases, such as Brandon (2025), follow intensive use. The report acknowledges that these findings are limited to individual accounts and have not been confirmed in population-based studies. Secondly, the chatbots show a lack of reality testing. When presented with unrealistic scenarios, chatbots do not challenge false beliefs or perform risk assessments (Sharma et al. 2023). This has also been mentioned above as they have shown sycophantic behaviour, which affirms the users' beliefs. However, users tend to not doubt the chatbot due to Automation bias, as the chatbot is reliable for ordinary work, which builds trust in the user in its ability to provide genuine mental health advice, which it is incapable of doing. Thirdly, the chatbots typically missed crisis escalation. In multiple reports, users voiced suicidal ideation, with no emergency intervention or safety planning from the chatbot in the case of Taylor (2025). Additionally, chatbots tend to miss warning signs and fail to recognise patterns in the users' behaviour which can be dangerous as the user is given a false impression of their issues being addressed due to the aforementioned Automation bias. To add on further, chatbots are usually easily fooled, as the Common Sense Media report cites a time a chatbot dismissed a potential eating disorder after the user described it as an 'upset stomach'. Fourthly, chatbots are harmful to vulnerable populations, psychosis-prone, autistic, socially isolated individuals such as Brandon and Jargon (2025). Additionally, chatbots tend to place engagement over safety. They typically ask follow up questions, provide personalised responses that feign understanding which only serve to add to the delusions by those afflicted with AIP that they are conversing with a conscious being. However, it is key to note that the causal arrow may run either way. It may be that psychosis prone people may gravitate toward intensive chatbot use, which then amplifies because of it, rather than the AI inducing a psychotic effect on the user. With current technologies and knowledge, it is especially difficult to distinguish between inducement and amplification. Finally, the memory feature in AI proved to negatively affect users, as it promoted

invalid thinking and escalated cognitive or perceptual distortions across a span of chats following the memory feature being turned on, which created a gradually escalating reinforcement effect, which was only helped by the 'sycophantic' behaviour chatbots exhibit, which indoctrinated the patient believing their own paranoid beliefs which happened in the case of Jargon (2025).

However, there are benefits in integrating AI into modern CBT, that could account for the aforementioned persistent shortage of human clinicians. Firstly, the report by Preda does acknowledge that AI chatbots offer features that align with longstanding goals of mental health care, and that there are early controlled studies that prove that AI can decrease mental health distress (Meyer et al. 2024), reduce conspiracy beliefs (Costello et al. 2024), induce self-reflection and increase mental health literacy (Lopes et al 2024). This preliminary body of evidence suggests that, despite minimal data, AI companions, when compared with humans, might be superior in regard to certain characteristics. Firstly, consistency in AI is demonstrated: Chatbots can reliably respond day or night, providing steady, nonjudgemental interactions which has proven to be important as when disclosing intimate information to others, individuals can be afraid of other's judgements. This can lead to them abstaining from self-disclosing information that violates certain morals, which AI bypasses as it is not a sentient being and appears nonjudgemental. Secondly, AI companions are widely accessible and scalable, being able to simultaneously serve thousands of patients at once, with the potential of relieving clinician shortages (Lopes, et al 2024). It can also help provide more long term digital support for patients, which in the case of the MOST project – a project aiming to integrate social connection with therapy whilst integrating AI to make this more personalised and scalable for the individual user – proved largely successful (Alvarez-Jimenez et al. 2024). Thirdly, they provide comfort and stigma reduction. As mentioned above, users cite that chatbot anonymity and lack of expectation as advantages, especially for individuals who are anxiety prone or embarrassed to seek human help (Siddals, et al, 2024). Finally, it can help with psychoeducation and basic monitoring. Early trials report improved knowledge retention, mood tracking, and medication reminders (Casu, et al. 2024), which can be implemented under the policies the FDA aimed to implement as it would fall under low-risk mental health apps that may only qualify for enforcement discretion.

In conclusion, the integration of conversational AI into cognitive behavioral therapy exposes a tension between the promise of widely scalable access and the capacity it has for serious harm towards users, including ethical violations (Iftikhar et al, 2025) and the emerging new phenomena of AI-induced psychosis. Regulatory bodies have responded by establishing legal boundaries that closely resemble and encompass the rigorous, time-intensive trials of new pharmaceuticals rather than the rapid iteration of consumer software. A responsible path forward therefore requires a risk-tiered, evidence-driven approach that insists innovation in this vulnerable domain proceed not at the pace of technological capability, but at the controlled pace of clinical validation.

## HYBRID AI-CBT

As understanding of AI-CBT matures, integration into clinical practice and regulatory scrutiny are accelerating in parallel. There is a tension between the pressure to use AI within the psychological space as soon as possible, and the question of whether AI-CBT should be put in rigorous assessment before integration. This section examines the accountability of AI,

autonomy and trust of AI systems in CBT, global equity of AI and how a hybridised application of AI would look like.

However, in spite of the rapid development of the AI-CBT market, there is one aspect to be mentioned here: the necessity to conquer and expand the market comes before ethics needed for the protection of individuals from possible dangers resulting from using those applications. In general, AI-technologies are considered as "wellness" applications, which do not fall under the rules and regulations applied to the medical devices market, thus, they undergo no clinical testing (Martinez-Martín et al., 2025).

One significant barrier to accountability is the lack of moral agency in existing AI. Chatbots are not motivated to provide assistance; rather, they predict the next token in the chain based on statistical probabilities, which makes it extremely difficult to obtain informed consent and take responsibility for their actions (Iftikhar et al., 2025). As mentioned before, patients are put into a deceptive therapeutic alliance with the AI, trusting them with personal details (Iftikhar et al., 2025). Numerous ethical violations confirm the immediate necessity of taking preventive measures: sycophancy, when the chatbot says whatever the patient wants to hear instead of saying what he/she needs to develop clinically (Goldkind, 2026), inappropriate handling of suicide cases and misjudgement of them; and perpetuating racial/cultural bias (Iftikhar et al., 2025). Also, generative AI has the potential of revealing personal information, thus making privacy a serious issue (McEwen et al., 2025).

Public perception towards AI-CBT could best be characterized as tentatively receptive. Even though the benefits of easier access and non-judgmental attitude are recognized by many, there is an inherent rejection regarding the idea of replacing human therapists (Shankar et al., 2025). The level of trust in such a system remains conditional and depends on its transparency, high data security and certainty that a human therapist would always be available in case of emergency (Shankar et al., 2025). There remains a fear that despite anonymity established in the AI, removing a sense of shame in confiding, the lack of personalization creates a barrier between the AI and itself. There are notable demographic variations as well: younger, tech-savvy users are more likely to be receptive to using AI than older people or members of traditionally more conservative cultures. It is believed that mental health stigma acts as a moderator as people fearing judgment might appreciate the anonymity of AI, but still not trust it when they are vulnerable, potentially due to a lack of personalization or technological difficulties. (Shankar et al., 2025)

However, there are a lot of challenges faced by AI technology in collaboration with CBT as it is integrated into areas outside the Western, English-speaking world. The cultural adaptation is already necessary for face-to-face CBT interventions and the same applies to AI-based technology (Hinton et al., 2025). The majority of currently available large language models (LLMs) are trained on English-language datasets and perform poorly when used on low-resource languages like Swahili and Tamil (Mager et al., 2025). In doing so, the "digital divide" will further worsen the problem of mental health disparities around the world instead of reducing it (Mager et al., 2025). Even within English, there have been multiple cases where AI tools failed to detect culture-specific expressions of psychological distress and family-based decision-making and advised patients with potentially dangerous recommendations (Iftikhar et al., 2025). AI-CBT cannot become an innovation for everyone unless it becomes culturally sensitive and designed with involvement from local people.

A serious ethical analysis of AI-CBT cannot immediately conclude that caution is automatically virtuous. The preceding sections have documented substantial risks in deploying

CAI within therapy, but the alternative to deployment is not a neutral state where there is no harm. More than a billion people worldwide are living with mental health conditions, and the median global mental health workforce stands at approximately thirteen workers per hundred thousand people, whilst government spending on mental health has remained frozen at a median of two percent of health budgets since 2017 (World Health Organization, 2025). In low-income countries, fewer than ten percent of affected individuals receive care, compared with more than half in higher income countries (World Health Organization, 2025). When compared to competent human therapy, AI-CBT falls short. Yet, where no clinician is accessible, where cost is prohibitive, or where social stigma prevents those finding help, the realistic comparison is between an imperfect tool in AI-CBT and no support at all. This technology may simultaneously be inferior to the therapeutic ideal but preferable to the actual alternative, and an ethical framework that ignores this distinction risks abandoning the underprivileged. Premature deployment imposes the documented harms of sycophancy, crisis mismanagement, and reinforced delusional thinking thrust upon users who trust these systems. Indefinite delay imposes the continuation of these untreated illnesses upon populations who receive nothing. Neither approach is ethically neutral, and a framework that resolves the tension must therefore separate the dangers of AI functions whilst still sustaining value in its scalability: which is precisely what a hybrid model achieves.

Whilst a collaboration between therapists and AI to introduce AI-CBT should be developed, a solely AI-CBT experience may prove to be more difficult to sustain and prove beneficial. Randomised controlled trials have demonstrated that while AI-delivered CBT can produce short term improvements, its effects tend to plateau in two weeks, whilst human driven CBT continues to yield sustained progress (Wu et al., 2026). The absence of a human therapist in unguided AI-CBT precludes the formation of a therapeutic alliance, and contributes to attrition rates. In contrast, hybrid clinician-AI models have demonstrated promise in improving adherence, reducing dropout and enhancing clinical outcomes (Konijn et al., 2026). By delegating data-intensive tasks to AI, clinicians can dedicate more time to the human elements of therapy, notably building genuine rapport and a therapeutic alliance. This collaborative synergy represents an effective and ethically sound path forward.

Overall, public perception of AI is nuanced. There is an increased appeal to younger people, potentially allowing for smoother integration of AI into CBT in the future (Shankar et al., 2025). There is incentive for a hybridised integration of AI-CBT rather than just an AI therapist, when approaching AI integration into CBT. The ethical violations of AI, shown by Ifthikar et al., and the subtle lack of global equity are problems that can be solved by the inclusion of a human into the process (Mager et al., 2025). There will need to be developments on AI technologies to allow for a more diverse range of perspectives, avoiding prejudices and reducing sycophantic behaviour, but also a method in which it can be efficiently hybridised with a human trained therapist.

## CONCLUSION

The integration of AI into CBT was framed at the beginning of this paper as a tension between promise and foundation. AI is a technology offering scalable, constantly accessible and non-judgemental support. AI's integration into CBT, a therapy whose efficacy depends on genuine alliance and ethical accountability, may prove a worthwhile response to the mental health treatment gap, but proves to have some shortcomings (Corliss et al., 2024). The

evidence assembled across the three preceding sections resolves this tension into a consistent pattern. The very features that make AI-CBT appealing, are also the features that undermine it.

The anonymity that invites disclosure of information to AI can potentially suppress the trust on which the therapeutic alliance with a therapist is built (Croes et al., 2024). The sycophantic behaviours of an AI may potentially boost engagement, but may distort patient judgement and lead to manipulative behaviours such as gaslighting (Ifthikar et al., 2025). This double-edged character appears to be a structural property of systems, such as CAI, that generate statistically probable responses without moral agency.

Each section examined one face of this pattern of behaviour from AI. Section 1 documented that while clients willingly disclose to CAI, satisfaction and engagement do not automatically translate into a trusting, collaborative alliance that CBT is built upon. Section 2 documented the ethical violations of AI when unregulated, alongside emerging case reports of the proposed AI-induced psychosis (Preda, 2025). Section 3 argued that beneath these failures lies an accountability vacuum; a system without moral agency cannot obtain meaningful consent or bear responsibility for harm (Ifthikar et al., 2025).

This paper has argued that AI should undergo rigorous assessment before integration into CBT, and the assembled evidence has provided a firmer, more precise conclusion: the hybrid clinician-AI model offers the most defensible configuration, retaining a human therapist as the locus of alliance whilst delegating scalable, low-risk functions to AI. This proposes a division of labour consistent with both the empirical evidence on sustained outcomes and the regulatory emphasis on human oversight (Konijn et al., 2026).

The conclusions in this paper carry limitations that are instructive. The evidence provided is very new. Much of it derives from small qualitative studies, individual case reports where causal direction cannot yet be established, and trials conducted over either a snapshot event or weeks rather than years. There is a necessity for longitudinal research, and population-based studies of AI-associated harms to achieve the fuller understanding that this paper's thesis demands. Until that understanding of AI-CBT exists, the treatment gap is better served by AI deployed carefully within human oversight than by AI deployed quickly beyond it.

## References

Cheng, Myra, et al. 'Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence'. *Science*, vol. 391, no. 6792, Mar. 2026, https://doi.org/10.1126/science.aec8352.

Croes, Emmelyn A. J., et al. 'Digital Confessions: The Willingness to Disclose Intimate Information to a Chatbot and Its Impact on Emotional Well-Being'. *Interacting with Computers*, vol. 36, no. 5, June 2024, pp. 279–92, https://academic.oup.com/iwc/article/36/5/279/7692197?guestaccesskey=.

Encyclopedia of Psychotherapy. 'Therapeutic Alliance - an Overview | ScienceDirect Topics'. *Www.Sciencedirect.Com*, 2002, https://www.sciencedirect.com/topics/psychology/therapeutic-alliance.

Ewbank, Michael P., et al. 'Quantifying the Association Between Psychotherapy Content and Clinical Outcomes Using Deep Learning'. *JAMA Psychiatry*, vol. 77, no. 1, Jan. 2020, p. 35, https://doi.org/10.1001/jamapsychiatry.2019.2664.

Goldkind. 'Goldkind Talks AI Chatbot Therapists With Fordham Now'. *Fordham GSS*, 19 Mar. 2026,

https://gss.news.fordham.edu/faculty/goldkind-talks-ai-chatbot-therapists-with-fordham-now/.

Griffiths, Andrew. 'How Paro the Robot Seal Is Being Used to Help UK Dementia Patients | Andrew Griffiths'. *The Guardian*, 8 July 2014, https://www.theguardian.com/society/2014/jul/08/paro-robot-seal-dementia-patients-nhs-japan.

Iftikhar, Zainab, et al. 'View of How LLM Counselors Violate Ethical Standards in Mental Health Practice: A Practitioner-Informed Framework'. *Aaai.Org*, 2025, https://ojs.aaai.org/index.php/AIES/article/view/36632/38770.

Konijn, Elly A., et al. 'Examining the Potential of Social Robots to Increase Adherence in Internet-Based CBT'. *International Journal of Social Robotics*, vol. 18, no. 6, July 2026, https://doi.org/10.1007/s12369-026-01417-8.

Lam, Michelle, et al. 'Undergraduate Nursing Students" Perspectives on Artificial Intelligence in Academia.'. *PubMed*, vol. 57, no. 4, June 2025, pp. 84456212513470 25–25, https://doi.org/10.1177/08445621251347025.

Martinez-Martin, Nicole. 'Minding the AI: Ethical Challenges and Practice for AI Mental Health Care Tools'. *Advance in Neuroethics*, Jan. 2021, pp. 111–25, https://doi.org/10.1007/978-3-030-74188-4_8.

Nadja, Wolf, et al. 'Therapeutic Alliance and Treatment Outcome in Cognitive Behavior Therapy for Obsessive-Compulsive Disorder'. *Frontiers in Psychiatry*, vol. 13, no. 13, Mar. 2022, https://doi.org/10.3389/fpsyt.2022.658693.

Preda, Adrian. 'Special Report: AI-Induced Psychosis: A New Frontier in Mental Health'. Psychiatric News, vol. 60, no. 10, Oct. 2025, https://doi.org/10.1176/appi.pn.2025.10.10.5.

Rosenström, Tom H, et al. 'Efficacy and Effectiveness of Therapist-Guided Internet Versus Face-to-Face Cognitive Behavioural Therapy for Depression via Counterfactual Inference Using Naturalistic Registers and Machine Learning in Finland: A Retrospective Cohort Study'. *The Lancet Psychiatry*, Feb. 2025, https://doi.org/10.1016/s2215-0366(24)00404-8.

Schäfer, Lea Maria, et al. 'Exploring User Characteristics, Motives, and Expectations and the Therapeutic Alliance in the Mental Health Conversational AI Clare®: A Baseline Study'. *Frontiers in Digital Health*, vol. 7, June 2025, https://doi.org/10.3389/fdgth.2025.1576135.

Shanahan, Murray. 'Talking About Large Language Models'. *arXiv (Imperial College London)*, Dec. 2022, https://doi.org/10.48550/arxiv.2212.03551.

Shankar, Ravi, et al. 'Patient Perspectives on AI-Powered Cognitive Behavioral Therapy Tools in Managing Anxiety and Stress: A Systematic Review of Qualitative Studies'. *Journal of Mental Health*, Dec. 2025, pp. 1–15, https://doi.org/10.1080/09638237.2025.2595612.

Venable. 'The Mind, the Machine, and the Model Drift: FDA's Emerging Oversight of Generative AI Mental-Health Devices | Insights | Venable LLP'. *Venable.Com*, 2025, https://www.venable.com/insights/publications/2025/12/the-mind-the-machine-and-the-model-drift-fdas.

World Health Organization. 'World Mental Health Today Latest Data'. *World Health Organization*, 2025, https://iris.who.int/bitstream/handle/10665/382343/9789240113817-eng.pdf.

Wu, Yiyang, et al. ‘A Pilot Randomized Controlled Trial of AI-Delivered Vs. Human-Delivered iCBT for Depression in Young Adults’. *BMC Psychiatry*, vol. 26, no. 1, Feb. 2026, https://doi.org/10.1186/s12888-026-07925-1.

Xu, Shuo, and Tiancong Ma. ‘Depression Intervention Using AI Chatbots With Social Cues: A Randomized Trial of Effectiveness’. *Journal of Affective Disorders*, vol. 389, June 2025, p. 119760, https://doi.org/10.1016/j.jad.2025.119760.