A Critical Discussion of Strategies and Ramifications of Implementing Conversational Agents in Mental Healthcare

Computers in Human Behavior: Artificial Humans 5 (2025) 100182

Contents lists available at ScienceDirect

Computers in Human Behavior: Artificial Humans

journal homepage: www.journals.elsevier.com/computers-in-human-behavior-artificial-humans

A critical discussion of strategies and ramifications of implementing conversational agents in mental healthcare

Arthur Bran Herbener [a,*], Michał Klincewicz [b,c], Lily Frank [d], Malene Flensborg Damholdt [a]

[a] Department of Psychology and Behavioral Sciences, Aarhus University, Denmark
[b] Cognitive Science and Artificial Intelligence, Tilburg University, the Netherlands
[c] Cognitive Science, Jagiellonian University, Poland
[d] Industrial Engineering and Innovation Science, Philosophy and Ethics, Eindhoven University of Technology, the Netherlands

ARTICLE INFO

Keywords:
Mental health
Psychotherapy
Care
Artificial intelligence
Chatbot
Conversational agent

ABSTRACT

In recent years, there has been growing optimism about the potential of conversational agents, such as chatbots and social robots, in mental healthcare. Their scalability offers a promising solution to some of the key limitations of the dominant model of treatment in Western countries. However, while recent experimental research provides grounds for cautious optimism, the integration of conversational agents into mental healthcare raises significant clinical and ethical challenges, particularly concerning the partial or full replacement of human practitioners. Overall, this theoretical paper examines the clinical and ethical implications of deploying conversational agents in mental health services as *partial* and *full* replacement of human practitioners. On the one hand, we outline how these agents can circumvent core treatment barriers through stepped care, blended care, and a personalized medicine approach. On the other hand, we argue that the partial and full substitution of human practitioners can have profound consequences for the ethical landscape of mental healthcare, potentially undermining patients' rights and safety. By making this argument, this work extends prior literature by specifically considering how different levels of implementation of conversational agents in healthcare present both opportunities and risks. We argue for the urgent need to establish regulatory frameworks to ensure that the integration of conversational agents into mental healthcare is both safe and ethically sound.

1. Introduction

Novel conversational agents powered by large language models have been portrayed as potential game-changers in mental healthcare (Filippis & Foysal, 2024; Javaid et al., 2023; Stade et al., 2024; Vowels, 2024). Conversational agents (CAs), such as chatbots, social robots, and digital avatars, are software systems designed to engage in verbal conversations with humans (Schöbel et al., 2023; Wahde & Virgolin, 2022). One of the frequently proposed applications for CAs is the delivery of evidence-based psychotherapy. Accordingly, there has been an upsurge in research investigating their feasibility as platforms for therapeutic intervention. Across dozens of randomized controlled trials, CAs have shown promising effectiveness in reducing symptoms of depression, anxiety, and psychological distress (for a meta-analytical review, see He et al., 2023). Additionally, large language models enable CAs to engage in human-sounding and contextually relevant conversations across an (almost) limitless array of topics, as demonstrated by the well-known chatbot ChatGPT (Schöbel et al., 2023; Shahriar & Hayawi, 2023). A growing body of research is inquiring about large language models' capabilities to deliver mental health interventions. Notably, multiple studies indicate that when CAs driven by large language models are instructed to assume the role as therapists, they are generally evaluated as more empathetic and helpful than human practitioners (Howe et al., 2023; Kuhail, Alturki, Thomas, & Alkhalifa, 2024; Kuhail, Alturki, Thomas, Alkhalifa, et al., 2024; Ovsyannikova et al., 2025; Vowels, 2024). Although clinical experimental research remains in its nascency, preliminary evidence further suggests that cognitive behavioral therapy (CBT) delivered by such CAs have clinically relevant effects on symptoms of depression and anxiety (Heinz et al., 2025). Intriguingly, the observed effect sizes were comparable to those typically reported for face-to-face CBT (Cuijper et al., 2017). While future research will determine whether these results represent a fortunate anomaly or a

* Corresponding author.
E-mail address: abh@psy.au.dk (A.B. Herbener).

https://doi.org/10.1016/j.chbah.2025.100182
Received 5 September 2024; Received in revised form 27 June 2025; Accepted 5 July 2025
Available online 8 July 2025

replicable outcome, they nonetheless mark a promising beginning for conversational agents in mental healthcare.

As the demand for mental health care has risen in recent decades, these technological advancements seem to have arrived at the perfect time. Epidemiological evidence shows a rising incidence of mental disorders in the Western world (Castelpietra et al., 2022; Dykxhoorn et al., 2023; Lipson et al., 2022; Twenge et al., 2019; WHO, 2013) and a major treatment gap globally, where only 8 % of individuals with depression in low-income countries and 33 % in high-income countries receive care from authorized mental health professionals. (Moitra et al., 2022). Various factors contribute to the treatment gap across individuals, nations, and cultures. A core antecedent of the treatment gap appears to be a limited *demand* for care, as individuals, particularly in low-income countries, do not necessarily perceive a need for care despite having symptoms of mental health conditions (Andrade et al., 2014; Roberts et al., 2022). However, especially in Western countries, one piece of the puzzle could also be the dominant model of treatment, i.e., one-to-one sessions undertaken by highly educated and well-paid mental health practitioners (Kazdin, 2018). Despite its widespread use and empirical support, this model is intrinsically linked to key obstacles to seeking care, including financial costs, lengthy waiting periods, geographical distances, and stigma (Kazdin, 2018).

Mental health conditions impose a substantial economic burden globally, with costs estimated at approximately $2.5 trillion annually in 2010—due to both poor health outcomes and reduced productivity—and projected to rise to $6 trillion by 2030 (Bloom et al., 2012). Crucially, some costs persist even when individuals access care, as mental health treatment itself entails significant direct expenses, including clinician fees, supervision, and administrative operations, as well as indirect costs borne by patients, such as transportation, childcare, and lost income. In addition to these quantifiable costs are the psychological burdens and distress that individuals in need of mental health care experience.

The disparity between mental health needs and available resources is particularly pronounced in low to medium income countries. A cross-sectional study of 30 countries in the WHO Region of the Americas found a striking imbalance between the burden of mental, neurological, and substance use disorders and corresponding government spending. The ratio of disease burden to mental health expenditure ranged from 3:1 in high-income countries like Canada and the USA to 435:1 in low-income countries such as Haiti, with a median ratio of 32:1 across the region (Vigo et al., 2019). Together, these factors highlight that receiving mental health support, while essential, remains economically demanding for individuals and society alike. Nevertheless, investment in mental health care yields significant economic returns; for every U.S. dollar invested in the treatment of depression and anxiety, there is an estimated return of four dollars in improved health outcomes and enhanced productivity (Chisholm et al., 2016). Thus, despite evidence of long-term effects, financial constraints and systemic limitations continue to impede timely access to mental health care, with particularly pronounced effects on underserved and rural populations. In this context, scalable digital interventions – including conversational agents – represent a promising avenue for achieving greater value for investment, by expanding reach and optimizing the use of limited resources within overburdened systems.

While direct evidence on the cost-effectiveness of CA-delivered therapy is still emerging, insights can be drawn from studies on internet-delivered therapy which has offered similar clinical outcomes as conventional therapy, for at least some diagnosis with reduced costs (Aspvall et al., 2021; Gentili et al., 2022; Rohrbach et al., 2023) although often with therapist involvement. The intersection of technological progress and societal demands for scalable mental health solutions has prompted researchers to view CAs as a game-changing tool to help address the treatment gap (Molli, 2022; Gallegos et al. 2023). A common perspective in the scientific literature is that CAs' scalability and inability to confer social judgments enables them to decrease the strain on overburdened healthcare systems (Pham et al., 2022; Sweeney et al., 2021; Boucher et al. 2011). While recent meta-analytical evidence indeed provides reason for cautious optimism (He et al., 2023), there are significant clinical and normative issues associated with replacing human practitioners with CAs, as we will show.

The present theoretical paper contributes to the ongoing scientific efforts to understand the potential and limitations of implementing CAs in mental healthcare. We examine how CAs can increase access to mental healthcare, while at the same time critically discussing which clinical and ethical ramifications the implementation may pose. We first explain how CAs can expand access to care through three treatment delivery approaches: 1) stepped care, 2) blended care, and 3) a personalized approach targeting fear of stigma. We argue that utilizing CAs in these strategies for implementation can help address core limitations of the dominant model of treatment. We then discuss the possible consequences of full or partial replacement of human practitioners in terms of practitioners' ethical and professional obligations and, ultimately, patients' trust in healthcare systems. With this theoretical paper, our goal is not to provide an exhaustive review of the scientific literature and provide actionable recommendations on that basis. Rather, primarily based on theoretical analysis, our aim is to raise awareness of potential actionable implementation strategies and to propose preliminary hypotheses about their clinical and ethical ramifications to help guide future research.

2. How conversational agents can increase access to care

2.1. Stepped and blended care

The feasibility of the dominant model of treatment in Western countries is highly contingent on the availability of mental health care providers (Kazdin, 2018). When the incidence of mental health problems is rising without a proportional increase in the number of mental health care providers, perfect conditions are created for a widening treatment gap. The main distinction in this context is between *stepped* and *blended* care. A stepped-care approach to treatment can be understood as assigning patients to the *most effective* and *least restrictive* treatment available in a *self-correcting* way. This means that patients are assigned to treatment based on their specific needs and severity of symptomatology starting with the least intensive interventions and -if ineffective-stepping up through more specialized treatment as needed in a self-correcting manner (Bower & Gilbody, 2005). Given that the least resource-intensive treatment is used, stepped care is cost-effective (Ho et al., 2016) and believed to reduce use of medication for mental health issues. Since the inception of stepped care models, several clinical guidelines have adopted the approach (e.g. NICE guidelines and AUS PHN guidelines). For a generic description of stepped care, see Table 1.

According to Bower and Gilbody (2005) the stepped care model rests upon three assumptions: i) that minimal interventions can provide a significant health gain, ii) that using minimal interventions will allow for more efficient use of resources, and iii) that the approach is acceptable to both health care professionals and patients. Importantly, if CAs are introduced into stepped care, then additional assumptions are introduced (e.g. that minimal interventions delivered by CAs can provide health gain etc.) which should be borne in mind when exploring if and where this technology could be used.

CAs could expand the stepped care model to include a "step 0" where AI via digital phenotyping of individual users (i.e. harvesting information across digital platforms) can identify users with unacknowledged mental health issues. CAs could then be designed to nudge users so identified to seek help. Digital phenotyping may also aid diagnosis of mental health issues (Onnela & Rauch, 2016) and "just-in-time interventions" (Torous et al., 2019) may also play a role, but research remains sparse. Early diagnosis and pre-emptive interventions could prevent unnecessary suffering and costs but come with several privacy, ethical and practical concerns that should be addressed before

Table 1
An abbreviated description of stepped care based on the NICE guidelines' (2011) combined summaries for all mental disorders.

Symptom severity	Intervention
Step 3: Persistent or increased symptomatology that has not responded to a low-intensity intervention	Individual psychotherapy as appropriate for diagnosis. Drug treatment as appropriate for diagnosis. **All disorders**: Support groups, befriending, rehabilitation programmes, educational and employment support services; referral for further assessment and interventions.
Step 2: Persistent subthreshold symptoms or mild to moderate	Some disorders: computerized treatment, individual facilitated self-help. **All disorders:** Support groups, educational and employment support services; referral for further assessment and interventions.
Step 1: All disorders – known and suspected presentations of common mental health disorders.	**All disorders:** Identification, assessment, psychoeducation, active monitoring; referral for further assessment and interventions.

implementation (Birk & Samuel, 2022).

To the best of our knowledge CAs have not been systematically implemented into a stepped care model targeted at any specific mental health issue yet but several stand-alone "step 1", "step 2", and "step 3" applications hold promise whereof some have already been scientifically explored. Thus, CAs have for instance been used for the initial screening and identification of mental health issues such as attention deficit hyperactivity disorder (Håvik et al., 2019) and the diagnoses of depression via machine learning-based virtual humans (Philip et al., 2017). Furthermore, AI-based tools have been developed for monitoring symptomatology using speech samples as biomarkers (Low et al., 2020). Currently, it appears that such tools are not *conversational* in nature but rather detect the presence/absence of pre-determined criteria within a narrow scope. This approach is unlikely to overcome the challenge of heterogeneity and comorbidity of mental disorders. It can easily be envisioned, however, that soon CAs would be sophisticated enough to facilitate broader clinical screening and assessment in a flexible manner, assessing a broad spectrum of symptomatology. Nevertheless, caution needs to be taken as inaccurate diagnosis may cause distress or self-stigmatization (Foulkes & Andrews, 2023), which underlines the importance of having AI diagnostic tools integrated into an eco-system of care where health professionals are involved.

A growing body of research also suggests that CAs can effectively deliver evidence-based psychotherapeutic interventions (e.g., cognitive behavioral therapy) adapted to a digitalized format. Overall, this emerging research field points toward the idea that mental health interventions delivered by CAs can be utilized as "step 2" in the generic stepped care model. In a seminal study, Fitzpatrick et al. (2017) examined the effectiveness of a fully automated cognitive behavioral intervention delivered by the chatbot Woebot to college students with mild symptoms of depression or anxiety. After two weeks, participants reported a significant decrease in symptoms of depression ($d = 0.44$) compared to participants receiving access to a self-help book on depression ($N = 70$). Woebot has also been associated with overall positive user experiences and high adherence to treatment, with an average of 12 check-ins (*range* = 8–18) throughout the study period. Building on the pioneering work by Fitzpatrick et al. (2017), a multitude of studies have since demonstrated preliminary and promising results in using CAs to improve depression, anxiety, psychological distress, and well-being (for a meta-analytical review, see He et al. (2023)).

While CAs can increase access to care, by both being available in general and by lowering the impact of stigma, substituting humans introduces its own set of challenges. As we will discuss more in-depth in Section 3, these challenges include, but are not limited to, potentially diminished opportunities for continuously monitoring patients' symptoms and treatment adherence and the possible loss of the highly acknowledged benefits from a strong therapeutic relationship. One way to retain "human in loop" for mental health treatments while leveraging the scalability of CAs is to utilize the blended care approach. Blended care merges traditional face-to-face treatment with digitally delivered interventions, creating a hybrid model where the healthcare professional and patient jointly decide the distribution of the two elements (Wentzel et al., 2016). Given their flexibility, CAs could be used between sessions for monitoring symptoms and progress, provide patient-education and treatment plan follow-ups (Car et al., 2020; Dingler et al., 2021) or perform individual treatment sessions, thereby decreasing the burden on healthcare professionals, as needed.

2.2. *A personalized approach for fear of stigma*

The idea of tailoring treatments to fulfill the needs of specific populations falls within the scope of *personalized medicine*, which addresses the question of *what works for whom* rather than *whether* treatments work (Cohen et al., 2021; Huibers et al., 2021). This approach relies on the assumption that better treatment outcomes can be achieved by tailoring treatments based on observed interactions between treatment characteristics and patient characteristics, such as symptom severity, comorbidity, personality traits, intelligence, and socioeconomic factors (Cuijpers, 2017).

An important feature of CAs in context of personalized medicine is the potential to tailor CAs to specific patient groups, across all the dimensions mentioned, but also others, including age, individual acceptance of CAs, levels of digital literacy, and potentially many other vectors, relevant for treatment (Gross et al., 2021). This opens new vectors to develop and test effectiveness of CA-delivered treatments. First, design features of a CA not directly connected to their specific role in care, such as the interface, visual and auditory features, etc., can be developed with specific demographics in mind. The same CA 'backend' can then deliver treatment in a way that fits with a particular group better or be engineered to change its manner, patterns of expression, etc., to fit with a particular group. Second, CAs can incorporate scientific and practical evidence with a programming update, with a focus on a specific group or in general. Perhaps we can then expect CAs to deliver ever-better interventions at pace with clinical advances.

A key contributing factor to inadequate care seeking might be that it fails to meet the needs of those who fear stigma from others. Stigma refers to the devaluation, discrimination, or discreditation of someone based on their personal characteristics, for example based on the presence of mental illness (Goffman, 1963/2009; Knaak et al., 2017). When someone fears being stigmatized, they might refrain from seeking care to avoid being labelled and experience shame or embarrassment (Corrigan, 2004; Rüsch et al., 2005). Multiple studies have identified stigma as a key barrier to care-seeking, with stigma – including shame, fear of negative social judgments, and concerns about disclosure and confidentiality – as the fourth-ranking barrier out of ten to seeking care (Clement et al., 2015). Stigma not only deters care-seeking but may also compromise treatment adherence by discouraging patients from sharing embarrassing or shameful information with practitioners (Clement et al., 2015; Knaak et al., 2017). A related concern is the possible confounding role of fear of stigma in mental health assessments, as this fear might increase the risk that individuals withhold information about their symptoms.

One reason why stigma compromises care-seeking and recovery – and why CAs pose a viable alternative to the dominant model of treatment – might be rooted in human tendencies to attempt to shape how others perceive them. According to impression management theory, humans consciously or subconsciously strive to project and maintain positive impressions of themselves (Goffman 1956). To achieve this, they seek to control the information they share, avoiding personal details that could lead to an undesired impression (Goffman 1965).

Consequently, the interpersonal encounter could be the very factor that renders psychotherapy suboptimal for those who have extensive fear stigma. In contrast, a growing body of evidence indicates that people openly share sensitive information about themselves to CAs and consider them as "safe spaces" to disclose embarrassing or shameful personal details (Kang & Kang, 2023; Kim et al., 2022; Lucas et al., 2014; Pickard et al., 2016; Pickard & Roster, 2020; Sah & Peng, 2015). This might be due to the cognitive belief that computers lack the mental states necessary for making social judgments (Kim et al., 2022).

Besides the preliminary meta-analytical evidence suggesting that CA-psychotherapy might pose a viable alternative to conventional care approaches (He et al., 2023), evidence from several studies suggests that humans disclose more information about themselves to CAs than to fellow humans. Indeed, a seminal study found that more socially anxious participants shared more intimate details to CAs than human recipients (Kang & Gratch, 2010). Likewise, a study found that participants preferred to answer questions about highly sensitive topics (e.g., ethnicity, cheating, lying, and personal income) to an avatar compared to a human interviewer (Pickard et al., 2016). Some evidence also suggests that people feel less discomfort about self-disclosing to CAs than humans. Lucas et al. (2014) found that participants experienced less fear and need to manage impressions, as well as displayed more sadness when believing that they interact with a computer compared to a human. Finally, there is also evidence to suggest that some individuals deliberately use chatbots as "safe spaces". Several qualitative studies on users of the chatbot Replika have found that the main reasons for using it is the confidential environment it provides for self-disclosures and its empathetic, validating responses (Skjuve et al., 2021, 2022; Brandtzaeg, 2022). Thus, there already appears to be a need and willingness among some individuals to seek emotional support in CAs when in fear of stigma.

All this said, it is still important to note that the interpersonal encounter in the dominant model of treatment in Western countries may also serve a remedying function for those fearing stigma from others specifically. When one anticipates a negative reaction from others (e.g., stigmatization), but is met with empathy, warmth, and validation, an opportunity for a *corrective emotional experience* occurs. A corrective emotional experience is a sort of interpersonal learning experience that arises when others' reactions contrast with one's expectations (Alexander & French, 1946; Lundh, 2014). According to the theory, this experience challenges the cognitive belief that others, for instance, stigmatize based on mental illness. While originating in the psychoanalytic tradition (Alexander & French, 1946), several scientists consider corrective emotional experiences a common ingredient across therapeutic traditions as well as a learning experience arising in social interactions in general (Grencavage & Norcross, 1990; Hayes et al., 2012; Elliot et al., 2013). In conclusion, the interpersonal encounter in the dominant model of treatment holds the potential to decrease fear of stigma through corrective experiences – an opportunity that could be absent in human-CA interactions. A personalized approach seeking to reach those fearing stigma should therefore not necessarily aim to replace human practitioners entirely. Rather, such an approach could help patients manage the fear of stigma to encourage them to participate in conventional mental health care services. This is a point that seems often overlooked in contemporary debates about mental healthcare delivered by CAs.

3. How conversational agents can compromise trust in healthcare systems

Contemporary CAs are not simply affordable, scalable, and "non-judgmental" substitutes for human practitioners. Important clinical and normative challenges have already been underscored in the scientific literature. Sedlakova and Trachsel (2022) posit that CAs' ability to engage in natural-sounding conversations coupled with their inherent capacities as algorithm-driven machines positions them neither as fully moral agents nor simply as tools, but somewhere in between these poles. According to them, a consequence is that therapeutic conversations with CAs cannot foster the same therapeutic benefits as interhuman interactions and can even give rise to illusions about having a close, interpersonal connection (Sedlakova & Trachsel, 2022). Others have questioned whether such mental health treatments work through similar causal pathways as human-delivered interventions, particularly casting doubt on whether CAs can elicit a beneficial therapeutic alliance and foster expectations of improvements (Herbener et al., 2024). Other scientists have also challenged the idea that mental health treatments delivered by CAs pose an alternative to traditional (human-delivered) psychotherapy. According to Grodniewicz and Hohol (2024), CAs should rather be seen as cognitive-affective artifacts that primarily possess the capacity to evoke emotional improvements through engaging users in cognitive tasks, whereas the socioemotional benefits of engaging with human practitioners may not be replicated (Grodniewicz & Hohol, 2024). These could also include the physical and relational aspects of the clinical encounter, including expressions of empathy and recognition of "non-verbal cues" (Brown & Halpern, 2021).

It is apparent from this line of perspectives that there remains substantial skepticism regarding CAs' potential to replace human practitioners. The skepticism mainly reflects assumptions about the role of human ontological status in the therapeutic process, which CAs may not replicate due to their status as inanimate, algorithm-driven technologies. Another commonality in this line of perspectives is a focus on clinical outcomes. Yet as outlined below, it is also necessary to explore how deploying CAs in mental healthcare may influence parameters beyond clinical outcomes. Indeed, as we will argue here, CAs significantly alter the normative landscape of mental healthcare in ways that could compromise patients' rights, safety, and ultimately trust in healthcare systems. In this section, we discuss how CA-psychotherapy challenges practitioners' ethical obligations toward patients, whether utilized as replacements or tools. By doing so, we emphasize the necessity of establishing regulatory frameworks for when and how CAs could be integrated into contemporary mental health care services. We summarize our concerns and recommendations for future research and regulations in Table 2.

3.1. Professional and ethical obligations

CA-psychotherapy, like other digital mental health interventions, including those sometimes called 'mobile health' or 'mHealth' (Istepanian et al., 2007; Kay et al., 2011), requires normative analysis. A wide range of ethical issues has been identified in the literature on mHealth specifically, such as whether they support or undermine autonomy (Schmietow & Marckmann, 2019), the extent to which they empower patient-users (Nickel et al., 2023), and complications of "privacy, trust, accountability and responsibility, and bias" (Murphy et al., 2021, p. 7), among others. Some of these insights apply also to the specific case of CA-psychotherapy. We will flag this below when appropriate but pay special attention to issues that are distinct and entirely new, given the technological affordances for use (and abuse) of CA-psychotherapy.

The first important distinction to make about the normative context of using CA-therapy is whether it is used as a stand-alone replacement for human practitioners or a supplement, as in stepped- or -blended care as described above. Replacement poses significantly more ethical and legal challenges than supplementation, although supplement can still alter the scope and nature of obligations that burden human practitioners. Even putting this distinction to the side, there is an important set of common concerns, which come from the more general set of normative complications that arise from using AI-power technology in a mental health context. CA-psychotherapy inherits them simply because it often relies on large language models (LMM), which use artificial intelligence (AI), broadly construed. The first and from our perspective

Table 2
Ethical challenges and implications.

Delivery approach	Ethical concern	Description	Implications
Replacementrowhead	Violation of duty of care	Administering CA-psychotherapy as a replacement for conventional psychotherapy challenges practitioners' duty of care due to the preliminary nature of the evidence supporting the efficacy of these interventions.	CA-psychotherapy interventions must demonstrate clinically relevant effects for specific patient groups when compared to other available interventions and while balancing potential adverse effects.
	Responsibility gaps	As CAs driven by AI can act autonomously, there may arise confusion regarding responsibility in case of harm.	CA-psychotherapy app manufacturers must have a clearly stated policy, embedded in relevant legislation or case law designation responsibility in cases of harm.
	Displacement of duty to warn	Practitioners' duty to warn extends to CAs if patients verbalize intentions to harm themselves or others.	Further empirical and ethical research is needed to assess how the duty to warn can be executed in these digital environments.
Partial replacementrowhead	Decontextualized communication	Chat logs/other text content patients generate through CAs have limited epistemic value for clinicians because tone, body language, demeanor, etc. are excluded.	Clinicians should be cautious interpreting text transcriptions from patient-CA sessions.
	Displacement of authority	The introduction of CAs into the therapeutic context may change patients' attitudes towards clinicians in ways that are not beneficial.	Further empirical research is needed on the possibility of this phenomenon. Clinicians should also be aware of these possibilities and attentive to them.
	Surveillance	CAs on smartphones may create an experience of constant surveillance and medicalization for patients.	The possibility of experiences of surveillance should be openly facilitated by the clinician. Patients should have easy ways of opting out of certain forms of interaction with CAs.

most important is the place and potential displacement of trust (Klincewicz, 2023; Klincewicz & Frank, 2019; Hatherley, 2020; Asan et al., 2020; DeCamp & Tilburt, 2019; Siddal, Torous, & Coxon 2024). The displacement of trust in context of CA-psychotherapy comes with further complications, which themselves differ depending on whether we are talking about replacement or supplementation.

What is the general role of trust in medicine and how does psychotherapy fit into it on a standard model? Trust is fundamental in all health care systems (Rhodes, 2001, 2020) in the sense that it underlies all the principles that govern medical ethics, including the therapeutic alliance, which is so important in mental health interventions (Howgego et al., 2003; Wampold & Flückiger, 2023). CA-psychotherapy, in both the replacement and supplement versions, creates a situation in which what a patient has good reason to do in relation to the source of the intervention is significantly different from a situation that involves only a human therapist.

For example, a patient is not rational to trust that the CA cares about them getting better, in the way that a person may care or appear to care for them getting better. A CA does not care about anything. Secondly, CA-psychotherapy does not come with accumulated trustworthiness of a professional human therapist (Aref-Adib et al., 2019; Bernard et al., 2016) or a medical institution. Instead, the norms for patients concern their expectations of reliability, rather than trust. The institutions that legitimize any psychiatric intervention depend on trust from patients who do not thereby have to engage in means-utility calculations each time they seek medical advice—they can simply trust the institution to do what it purports to, namely, take care of them. With CA-psychotherapy, patients do not have any reason to assume the same is true.

Trust and trustworthiness in psychotherapy that include therapeutic applications of AI, should also be distinguished from "trustworthy AI" itself, a term frequently used in discussions of AI regulation and ethics guidelines, for example in the EU high level expert group on AI (European Union, 2019). When it comes to the justified expectations that a person can have in any technology, it makes more sense to talk about reliability, efficacy, and transparency, which does not carry with it the same emotional components and is, arguably, an entirely different attitude (see Alanezi, 2024; Yang et al., 2024). When it comes to these expectations of reliability and efficacy, Siddals et al. (2024) have found in a small study that users were skeptical about the appropriateness or accuracy of the advice they received from the generative AI chatbots, expressing concerns about "issues such as the tendency to output incorrect or fabricated content (to 'hallucinate')" (Siddals et al., 2024; Yang et al., 2024).

3.2. Privacy and confidentiality

Another common concern with CA-psychotherapy is with the privacy and confidentiality of patient data — this is a common concern across all wellness, health, etc., apps already mentioned. However, it should be noted that in the context of CA-psychotherapy data privacy can be affected positively as well, which is unique. A therapists' lapse in their professional judgment or failure to secure adequate information protection, despite their intentions and efforts, is not difficult to imagine. Similar things can happen with CA-psychotherapy apps and databases that store patient data outside the therapists' offices. On the other hand, robust measures that prevent hacking may effectively isolate sensitive patient information better than a therapist could. Given this, a unique set of norms governing data acquired and used by CA-psychotherapy apps must be developed to protect patients and therapists.

These broad ethical concerns intersect with the extant regulatory landscape for digital health, which has been evolving for at least a decade to ensure the safety, reliability, and data security of these technologies that exist at the intersection of commercial products, medicine/health care and digital technology (Hassanaly & Dufour, 2021). Some of these existing frameworks apply to the special case of CA-psychotherapy. According to a technical brief on the "Evaluation of Mental Health Mobile Applications" by Argaqal and Wilcox (2022), the gold standard mobile apps fall into the Software as a Medical Device (SaMD) category used by the FDA. They fall into one of the three classifications for approval: Class I, II, or III, distinguished by the level of risk to the patient, user and the kind of notification and FDA clearance required before putting the device on the market in the United States[1] The FDA may remove from the market any harmful Class II and Class III devices, and this includes apps that target psychiatric conditions.

It should be noted that (as of now, but likely to change soon) only five mental health apps on the market in the United States have FDA approval and can be considered SaMDs (Agarwal S et al., 2022). For non-SaMDs, the FDA is not involved at all. Some mental health apps have been approved as medical health intervention across the diverse regulatory landscape in the European Union, for example in Germany the app "Mindable: Soziale Phobie", which uses CBT to treat social anxiety and phobia, has CE designation and is marketed as a medical device https://www.mindable.health/media/social/rev003-12-2023-ifu.pdf). Another notable certified medical device is Deprexis, a treatment for depressive disorder (https://deprexis.com/). Their website claims that:

> Deprexis works like psychotherapy and is just as effective. In conversations that build on one another, the user is shown information

and exercises in a sensitive manner and is offered various response options. (https://de.deprexis.com/)

At the EU level the Medical Device Regulation (EU) 2017/745 (MDR) has a risk-based framework for standards that apply to anything that meets the definition of a medical device, including mental health apps. As van der Storm et al. (2023) explain, manufacturers are asked to place their device in the appropriate risk class. Depending on the risk class the manufacturer themselves or an oversight body can then grant the technology the designation Conformité Européenne (CE), which indicates conformity with the MDR (van der Storm et al., 2023). As in other contexts, whether something qualifies as a medical device depends, in part, on the claims that its creators make and communicate in advertising about its "intended purpose". Those efforts notwithstanding, there are some noteworthy lapses in regulatory oversight. General wellness apps, which may on the surface promote mental health benefits to their users, do so without any independent specialized body overseeing them resulting in poor compliance with basic standards (Dehling et al., 2015; Loosman, 2024; Martinengo et al., 2019).

3.3. *Concerns related to full replacement of human practitioners with CAs*

What may motivate the idea that replacement is a viable option is the socio-economic context within which CA-psychotherapy is developed, as outlined in the first sections of this paper. CA-psychotherapy may be a solution to those problems. The motivation against this idea is that the replacement does more than fill the role of a therapist. In this section, we will make the case that replacement with CA-psychotherapy is in fact an elimination of the therapeutic relationship altogether, at least if we consider the essential duty of care to the patient (grounded in trust, at least according to Rhodes (2023)). Nevertheless, there are contexts where CA-psychotherapy may be the best or only choice of intervention and should be preferred as they can be perceived as non-judgmental or less threatening (Alanezi, 2024; Batra et al., 2017; Lucas et al., 2014)—a point similarly made above. That said, our criticism of this approach and the ethical and clinical issues that arise from the use of CA-psychotherapy as replacement, lead us to the conclusion that it should be avoided, unless no other options are available, at least until more effective and nuanced regulations are in place.

3.3.1. *Duty of care*

The first concern relates to the duty of care of human practitioners, which is a central ethical commitment of medicine in general and of psychotherapy in particular. It is the obligation to act beneficently towards one's patients, for example, by expressing concern, respecting their autonomy when making decisions about the course of care, informing them about potential consequences of treatment, etc. This principle is not only widely acknowledged in the professional ethics literature—it is also reflected in professional guidelines (See for example the Code of Ethics from the American Psychological Association, Principle A of the EPPCC of the APA). Thus, by recommending a treatment that has limited evidence of efficacy, therapists may be failing to fulfill this duty. This would especially be true if, as evidence suggests, the relationship between patient and therapist is one of the most important active ingredients in therapy and evidence that CA may play that role is currently lacking (Herbener et al., 2024). A relationship with the therapist involves the setting of shared expectations for progress and a commitment to the therapeutic process, which is essential to the therapeutic alliance that leads to better outcomes. As Henson et al. (2019) observe, "the therapeutic relationship or alliance is fostered across all therapies by interest, empathy, understanding, and hopeful, positive expectations of therapy" (p. 268). Without the relationship, it is not clear that such an alliance can even come about.

3.3.2. *Responsibility gaps*

A second concern is that the introduction of CA-psychotherapy as a replacement could lead to responsibility gaps, which have similarly been identified in other health-AI domains (Danaher, 2016; Matthias, 2004; Sparrow, 2007). Responsibility gaps arise when, according to Konigs (2022), the following two conditions are met:

NO RESPONSIBILITY: An autonomous system carries out some action without anyone—neither the programmer, the manufacturer, the operator, nor the autonomous system itself—being responsible for this action and its consequences.

AUTONOMY: This absence of responsibility is due to the system's autonomy.

In our context, CA-psychotherapy "introduce[s] a new element of technical opacity and lack of explainability that makes it more difficult for individual persons to satisfy the traditional conditions for moral and legal culpability: intention, foreseeability, and control" (Santoni De Sio & Mecacci, 2021, p. 1058). CAs will likely satisfy both criteria in case of a bad outcome for a patient. First, LLMs are famously uninterpretable for laypeople and even engineers; they are essentially gargantuan networks of associations, meant to mirror semantic associations between lexical items found in human text data. Second, their operation can be unpredictable, difficult (or impossible) to explain, and they do not come with human-in-the-loop functionality (though that may be possible to implement). Given this, responsibility gaps would likely arise often, and reactive attitudes of blame and responsibility would be appropriate for a patient to have in cases when things go badly for them (for example a relapse in substance abuse disorder or major depression, suicide, etc.). Yet, there would not be any one individual or group to which responsibility can be attributed.

The situation is further complicated when we look at the specifics of practice and its consequences on ethical and legal norms that protect the patient (and provider) from hazard, harm, and financial loss. Responsibility (both legal and moral) in the domain of psychotherapy is already complicated by the multiplicity of normative contexts around the world and the general lack of consensus about its status as medicine. For example, in the United States, psychotherapy can be delivered by psychologists with PhDs, psychiatrists with MDs, licensed clinical social workers, or pastoral counselors, among others. In the EU, generally a degree in psychology is required and certification from country-specific organizations, but psychotherapy itself is not regulated at the EU level. In this context, adding CA-psychotherapists will complicate how criteria for administering psychotherapy are interpreted and what happens when things go badly for the patient (or provider) during treatment. In the US, patients have recourse in civil court to make a case for negligence or malpractice. In most of the EU malpractice litigation is far less common and compensation is limited. It is not at all clear how they could do this with an artificial agent.

3.3.3. *Duty to warn*

One of the obligations of therapists under their duty of care is to continually assess the extent to which (if at all) a patient poses a risk to themselves or others, suicidal or homicidal, or has violent intentions of other sorts, come about because of interactions with the CA (Kaser-Boyd, 2015; Simon, 2002). In some of these cases, it is permissible or even required morally and sometimes legally to breach patient confidentiality for the sake of protecting the patient, themselves or (more controversially) third parties (Steyn, 2006; Liégeois & Eneman, 2012; Jobes & O'Connor, 2009). Because of the very serious consequences of such a breach, these decisions require significant professional judgment based on training and often discussion with colleagues. The mere possibility that the CA-psychotherapy agent may record and share patient disclosures to fulfill this goal would likely undermine the effectiveness of the intervention, potentially having a chilling effect on what they disclose. Either way, without the ability to stop harm or with the ability to stop it, the therapeutic alliance could be undermined. Given all the above observations, we should assume that for the foreseeable future the CA-'therapist' will not be able to discharge the duty of care or associated obligations (See those outlined in APA Code, Principle A and Principle B and Principle E).

3.3.4. Vulnerability and acceptance

In the ethics of mental health care special attention is paid to the vulnerability of users of technology and patients, mirroring the broader concern with vulnerability in medical ethics (Mackenzie, 2014; Rogers et al., 2012; Ten Have, 2015; Vayena et al., 2018). People in need of mental health care experience both inherent and situational vulnerability (Rogers et al., 2012). Inherent vulnerabilities stem from their physical, emotional, and psychological illnesses and frailties, while situational vulnerabilities are a result of an individual's position in society and characteristics that can be mitigated through distribution of resources (Mackenzie, Rogers, & Dodds 2014).

In the case of CAs in mental health care, different user groups may experience different kinds of vulnerabilities and thus require different types of safeguarding. Particular attention must also be given to the usability and acceptance of CAs amongst particular user groups. For example, some elderly people may not have the digital literacy necessary to fully engage with CAs and this needs to be taken into account in the design process. There are several systematic approaches which attempt to ensure that the interests and values of users are taken into account in technology design, including value sensitive design (Friedman & Hendry, 2019), participatory design (Simonsen & Robertson, 2013), inclusive or universal design (Clarkson et al., 2013), and capability sensitive design (Jacobs, 2020; Oosterlaken, 2009). Although using these methods does not guarantee inclusivity or usability, they prioritize the participation of potential user groups in the design process and frontload consideration of user values, diversity, and vulnerabilities.

3.4. Concerns related to partial replacement of human practitioners with CAs

Using CA in conjunction with a human therapist as blended care allows for remote monitoring of symptoms and at home exercises or interventions. There are four sets of normative questions to raise in the context of blended care. The first question concerns the status of the information, chat logs, exercises, etc. that is generated by patient interaction with the CA. Records generated in conversations with a therapist, notes, etc. are guarded by doctor-patient confidentiality. These documents often complement one-on-one sessions with the therapist, facilitating conversation, and allowing for the monitoring of progress, and other process related record keeping (APA 6.01 and 6.02). Therapists are also obligated to keep records of their treatment plan, diagnosis, etc. Blended care with CA introduces questions about the norms that should govern information that the patient discloses via their use of the CA.

The first normative challenge here is that there are no guidelines for therapists specifically vis-à-vis information disclosed via interactions with a conversational agent in a blended care context. At the same time, the data collected through the CA is protected by more general regulations like the General Data Protection Regulation (GDPR)—at least in countries that have that legal regime. This is a broad concern that is not unique to the CA-psychotherapy context and outside the scope of this paper, so we set it aside.

The second set of questions comes from practical problems of interpretation of text. There is an increased possibility of false negatives when face-to-face (FTF) interaction is decreased through text-based interactions. As Karcher and Presser (2018, p. 10) observe:

> "if the client is in high distress but the psychotherapist is unable to judge this based on the client's self-report via text, the client may be at increased risk or may feel misunderstood. Furthermore, a psychotherapist may have difficulty judging whether a client is disclosive and following recommendations when assessing risk and the means of communication is electronic. Thus, psychotherapists need to take precautions when using mHealth in order to accurately assess the client's mental state during non-FTF interactions so that [APA] Standard 3.04 [Avoiding Harm] is not violated"

This and many other possible problems associated with mediating patient input through an app shed doubt on at least one of the purported advantages of using blended CA-based interventions: the promise of non-judgmental regard during the conversation and to encourage people who fear stigma to seek help. Even if they are encouraged to do so, we have no way of knowing whether they are indeed engaging in the sort of disclosure sought after in a therapeutic context and the mediation of therapy-related disclosures through technology may lead the therapist in the wrong direction during treatment.

The third set of questions arise when we conceive of the CA as a third party that enters the therapeutic relationship with the patient and doctor. Of course, it is already the case that many patients, especially those with complex diagnosis, are treated by teams of providers, social workers, therapists and psychiatrists for example. These providers have different sets of expertise, experience, and treatment modalities, and roles. So, one may wonder, what difference would adding a CA to this group make, generally speaking? If there is a CA involved, patients may also come to see it as another member of the care team whose approach or advice they prefer and trust more, and even take the recommendations given by the CA as more authoritative or objective. People sometimes treat automated systems and especially those powered by AI as more objective than people, who may be perceived as prone to bias (Choung et al., 2023; Claudy et al., 2022). This would potentially challenge the authority of the human team members and could derail the treatment plan without the human therapist knowing until it is too late. Another problem is that CA advice may simply be more palatable to the patient than therapist advice.

The fourth set of questions arise from the fact that introducing a CA into the therapist-patient relationship changes the extent to which the patient themselves would be rational to feel trusted. Conventional therapy often involves asking patients to incorporate specific activities (or abstain from others) into their daily lives, keep thought records, engage in mindfulness, exercise, boundary setting, etc. Engaging with a CA would be a part of these exercises or a replacement for some of them. But CA-blended care is different from most of the other "homework" that patients are assigned. It can keep an objective record of how much and how often the patient is engaged with it. Instead of relying on patient self-reports and reflections, therapists have the possibility of checking the meta-data on CA engagement, if not accessing the content itself. The patient would be rational to think of themselves as being under continual surveillance and in a constant state of treatment, having the CA function as a sort of mental health panopticon.

As CA-psychotherapy tools grow in technical sophistication, the possibilities for surveilling patients outside of the therapist's office will expand. This possibility can undermine the patient's sense that their therapist trusts the veracity of their self-reports, increasing their experience of vulnerability in a relationship in which power is already heavily skewed towards the therapist. There is extensive evidence on the many ways in which being in a state of feeling one is "being watched" can alter one's relationships, thoughts, and behaviors (Dear et al., 2019; Foucault, 1977). At the same time, frequent engagement with CA, or even the affordance for such engagement, via one's smartphone potentially expands the amount of time and the contexts within one conceives of oneself as a patient or someone struggling with their mental health.

In general, embedding CAs in the broader socio-technical systems of mental healthcare in diverse contexts is fraught and under-unregulated. In the United States, the Food and Drug Administration regulates the development and use of medical interventions, including devices, medicines, and therapies, broadly construed. The regulatory landscape is not yet fully formed and in some places around the world, non-existent. What this means is that the obligations of developers of the interventions and the obligations of cpractitioners have no obvious priority over each other in specific cases. In other words, it is not clear under which regulatory regime a particular mental health application finds itself before it even makes it into a therapeutic context.

Given that the topic of this paper is highly interdisciplinary and that technology and applications are rapidly evolving a directed narrative, an interpretive approach shaped by a theoretical orientation and diverse disciplinary backgrounds brings into focus that these issues are ultimately normative and demand careful reflection about values at both an institutional and societal level. It also makes clear where further research is needed, namely, in the normative dimension of different types of CAs and their features as they relate to specific patient groups, for example, patient age, gender, treatment history, and mental health condition.

4. Conclusion

Recent decades have seen an increasing incidence of mental health disorders, seemingly without a proportional effort to meet the amplified demand for care, considering the high treatment gaps worldwide. The dominant model of treatment in Western countries plays a substantial role in maintaining the treatment gap due to its reliance on the availability of mental health providers and the stigma associated with seeking care. CAs hold promise for helping bridge this gap through stepped and blended care approaches and by serving as "safe havens" for those who fear stigma from others. Yet the advantage of the accessibility and scalability of CAs is ample fuel for techno-solutionism, which obscures deeper socioeconomic problems and potentially fosters new clinical and ethical challenges. As we pointed out in this paper, CAs may transform the normative landscape of mental healthcare delivery in ways that, if unregulated, could compromise patients' trust in healthcare systems, broadly speaking, due to violations of practitioners' professional and ethical obligations and confusion about responsibility. Our recommendation is therefore that, until clinical efficacy has been established and regulatory frameworks are in place, CAs should be implemented with caution, not only for the patients' sake in the short term but also to preserve their trust in healthcare systems.

CRediT authorship contribution statement

Arthur Bran Herbener: Project administration, Writing – review & editing, Conceptualization, Writing – original draft. **Michał Klincewicz:** Conceptualization, Writing – review & editing, Supervision, Writing – original draft. **Lily Frank:** Conceptualization, Writing – original draft, Writing – review & editing. **Malene Flensborg Damholdt:** Writing – original draft, Conceptualization, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Agarwal, S., Jalan, M., Wilcox, H. C., Sharma, R., Hill, R., Pantalone, E., Thrul, J., Rainey, J. C., & Robinson, K. A. (2022). *Evaluation of mental health Mobile applications (report no. 22-EHC016)*. Agency for Healthcare Research and Quality (US). https://www.ahrq.gov/research/findings/evidence-based-reports/mhealth.html.

Alanezi, F. (2024). Assessing the effectiveness of ChatGPT in delivering mental health support: A qualitative study. *Journal of Multidisciplinary Healthcare*, 461–471.

Alexander, F., & French, T. M. (1946). *Psychoanalytic therapy: Principles and application*. Ronald Press.

Andrade, L. H., Alonso, J., Mneimneh, Z., Wells, J. E., Al-Hamzawi, A., Borges, G., Bromet, E., Bruffaerts, R., De Girolamo, G., De Graaf, R., Florescu, S., Gureje, O., Hinkov, H. R., Hu, C., Huang, Y., Hwang, I., Jin, R., Karam, E. G., Kovess-Masfety, V., … Kessler, R. C. (2014). Barriers to mental health treatment: Results from the WHO world mental health surveys. *Psychological Medicine, 44*(6), 1303–1317. https://doi.org/10.1017/s0033291713001943

Aref-Adib, G., McCloud, T., Ross, J., O'Hanlon, P., Appleton, V., Rowe, S., Murray, E., Johnson, S., & Lobban, F. (2019). Factors affecting implementation of digital health interventions for people with psychosis or bipolar disorder, and their family and friends: A systematic review. *The Lancet Psychiatry, 6*(3), 257–266. https://doi.org/10.1016/S2215-0366(18)30302-X

Asan, O., Bayrak, A. E., & Choudhury, A. (2020). Artificial intelligence and human trust in healthcare: Focus on clinicians. *Journal of Medical Internet Research, 22*(6), Article e15154. https://doi.org/10.2196/15154

Aspvall, K., Sampaio, F., Lenhard, F., Melin, K., Norlin, L., Serlachius, E., Mataix-Cols, D., & Andersson, E. (2021). Cost-effectiveness of internet-delivered vs In-Person cognitive behavioral therapy for children and adolescents with obsessive-compulsive disorder. *JAMA Network Open, 4*(7), Article e2118516. https://doi.org/10.1001/jamanetworkopen.2021.18516

Batra, S., Baker, R. A., Wang, T., Forma, F., DiBiasi, F., & Peters-Strickland, T. (2017). Digital health technology for use in patients with serious mental illness: A systematic review of the literature. *Medical devices (Auckland, N.Z.), 10*, 237–251. https://doi.org/10.2147/MDER.S144158

Bernard, R., Sabariego, C., & Cieza, A. (2016). Barriers and facilitation measures related to people with mental disorders when using the web: A systematic review. *Journal of Medical Internet Research, 18*(6), Article e157. https://doi.org/10.2196/jmir.5442

Birk, R. H., & Samuel, G. (2022). Digital phenotyping for mental health: Reviewing the challenges of using data to monitor and predict mental health problems. *Current Psychiatry Reports, 24*(10), 523–528. https://doi.org/10.1007/s11920-022-01358-9

Bloom, D. E., Cafiero, E., Jané-Llopis, E., Abrahams-Gessel, S., Bloom, L. R., Fathima, S., Feigl, A. B., Gaziano, T., Hamandi, A., & Mowafi, M. (2012). The global economic burden of noncommunicable diseases. *Program on the Global Demography of Aging*. https://ideas.repec.org/p/gdm/wpaper/8712.html.

Bower, P., & Gilbody, S. (2005). Stepped care in psychological therapies: Access, effectiveness and efficiency: Narrative literature review. *The British Journal of Psychiatry, 186*(1), 11–17. https://doi.org/10.1192/bjp.186.1.11

Brandtzaeg, P. B., Skjuve, M., & Følstad, A. (2022). My AI friend: How users of a social chatbot understand their Human–AI friendship. *Human Communication Research, 48*(3), 404–429. https://doi.org/10.1093/hcr/hqac008

Brown, J. E., & Halpern, J. (2021). AI chatbots cannot replace human interactions in the pursuit of more inclusive mental healthcare. *SSM-Mental Health, 1*, Article 100017.

Car, L. T., Dhinagaran, D. A., Kyaw, B. M., Kowatsch, T., Joty, S., Theng, Y.-L., & Atun, R. (2020). Conversational agents in health care: Scoping review and conceptual analysis. *Journal of Medical Internet Research, 22*(8), Article e17158. https://doi.org/10.2196/17158

Castelpietra, G., Knudsen, A. K. S., Agardh, E. E., Armocida, B., Beghi, M., Iburg, K. M., Logroscino, G., Ma, R., Starace, F., Steel, N., Addolorato, G., Andrei, C. L., Andrei, T., Ayuso-Mateos, J. L., Banach, M., Bärnighausen, T. W., Barone-Adesi, F., Bhagavathula, A. S., Carvalho, F., Carvalho, M., … Monasta, L. (2022). The burden of mental disorders, substance use disorders and self-harm among young people in Europe, 1990-2019: Findings from the global burden of disease study 2019. *The Lancet regional health. Europe, 16*, Article 100341. https://doi.org/10.1016/j.lanepe.2022.100341

Chisholm, D., Sweeny, K., Sheehan, P., Rasmussen, B., Smit, F., Cuijpers, P., & Saxena, S. (2016). Scaling-up treatment of depression and anxiety: A global return on investment analysis. *The Lancet Psychiatry, 3*(5), 415–424. https://doi.org/10.1016/S2215-0366(16)30024-4

Choung, H., Seberger, J. S., & David, P. (2023). When ai is perceived to be fairer than a human: Understanding perceptions of algorithmic decisions in a job application context. *International Journal of Human-Computer Interaction*. https://doi.org/10.1080/10447318.2023.2266244. Advance online publication.

Clarkson, P. J., Coleman, R., Keates, S., & Lebbon, C. (2013). *Inclusive design: Design for the whole population*.

Claudy, M. C., Aquino, K., & Graso, M. (2022). Artificial intelligence can't be charmed: The effects of impartiality on laypeople's algorithmic preferences. *Frontiers in Psychology, 13*, Article 898027. https://doi.org/10.3389/fpsyg.2022.898027

Clement, S., Schauman, O., Graham, T., Maggioni, F., Evans-Lacko, S., Bezborodovs, N., Morgan, C., Rüsch, N., Brown, J. S. L., & Thornicroft, G. (2015). What is the impact of mental health-related stigma on help-seeking? A systematic review of quantitative and qualitative studies. *Psychological Medicine, 45*(1), 11–27. https://doi.org/10.1017/s0033291714000129

Cohen, Z. D., Delgadillo, J., & DeRubeis, R. J. (2021). Personalized treatment approaches. In *Bergin and Garfield's handbook of psychotherapy and behavior change* (50th anniversary edition, 7th ed., pp. 673–703). John Wiley & Sons, Inc.

Corrigan, P. (2004). How stigma interferes with mental health care. *American Psychologist, 59*(7), 614–625. https://doi.org/10.1037/0003-066x.59.7.614

Cuijpers, P. (2017). Four decades of outcome research on psychotherapies for adult depression: An overview of a series of meta-analyses. *Canadian Psychology/Psychologie Canadienne, 58*, 7–19. https://doi.org/10.1037/cap0000096

Danaher, J. (2016). *Robots, law and the retribution gap* (Vol. 18). Ethics and Information Technology. https://doi.org/10.1007/s10676-016-9403-3

Dear, K., Dutton, K., & Fox, E. (2019). Do 'watching eyes' influence antisocial behavior? A systematic review & meta-analysis. *Evolution and Human Behavior, 40*(3), 269–280. https://doi.org/10.1016/j.evolhumbehav.2019.01.006

DeCamp, M., & Tilburt, J. C. (2019). Why we cannot trust artificial intelligence in medicine. *The Lancet Digital Health, 1*(8), Article e390. https://doi.org/10.1016/S2589-7500(19)30197-9

Dehling, T., Gao, F., Schneider, S., & Sunyaev, A. (2015). Exploring the far side of Mobile health: Information security and privacy of Mobile health apps on iOS and android. *JMIR mHealth and uHealth, 3*(1), e8. https://doi.org/10.2196/mhealth.3672

Dingler, T., Kwasnicka, D., Wei, J., Gong, E., & Oldenburg, B. (2021). The use and promise of conversational agents in digital health. *Yearbook of Medical Informatics, 30*(1), 191–199. https://doi.org/10.1055/s-0041-1726510

Dykxhoorn, J., Gnani, S., Kirkbride, J. B., Lazzarino, A. I., Osborn, D., & Walters, K. (2023). Temporal patterns in the recorded annual incidence of common mental disorders over two decades in the United Kingdom: A primary care cohort study. *Psychological Medicine*, 1–12. https://doi.org/10.1017/S0033291723002349

Filippis, R. D., & Foysal, A. A. (2024). Chatbots in psychology: Revolutionizing clinical support and mental health care. *Voice of the Publisher, 10*(3), 298–321. https://doi.org/10.4236/vp.2024.103025

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), Article e19. https://doi.org/10.2196/mental.7785

Foucault, M. (1977). *Discipline and punish: The birth of the prison*. New York: Vintage Books. ISBN 978-0-67975-255-4.

Foulkes, L., & Andrews, J. L. (2023). Are mental health awareness efforts contributing to the rise in reported mental health problems? A call to test the prevalence inflation hypothesis. *New Ideas in Psychology, 69*, 1–6. https://doi.org/10.1016/j.newideapsych.2023.101010

Friedman, B., & Hendry, D. G. (2019). *Value sensitive design: Shaping technology with moral imagination*. Mit Press.

Gentili, A., Failla, G., Melnyk, A., Puleo, V., Tanna, G. L. D., Ricciardi, W., & Cascini, F. (2022). The cost-effectiveness of digital health interventions: A systematic review of the literature. *Frontiers in Public Health, 10*. https://doi.org/10.3389/fpubh.2022.787135

Goffman, E. (1963/2009). *Stigma: Notes on the management of spoiled identity*. Simon and Schuster.

Grencavage, L. M., & Norcross, J. C. (1990). Where are the commonalities among the therapeutic common factors? *Professional Psychology: Research and Practice, 21*, 372–378. https://doi.org/10.1037/0735-7028.21.5.372

Grodniewicz, J., & Hohol, M. (2024). Therapeutic chatbots as cognitive-affective artifacts. *Topoi*. https://doi.org/10.1007/s11245-024-10018-x

Gross, C., Schachner, T., Hasl, A., Kohlbrenner, D., Clarenbach, C. F., Wangenheim, F. V., & Kowatsch, T. (2021). Personalization of conversational agent-patient interaction styles for chronic disease management: Two consecutive cross-sectional questionnaire studies. *Journal of Medical Internet Research, 23*(5), Article e26643.

Hassanaly, P., & Dufour, J. C. (2021). Analysis of the regulatory, legal, and medical conditions for the prescription of Mobile health applications in the United States, the european union, and France. *Medical Devices: Evidence and Research, 14*, 389–409. https://doi.org/10.2147/MDER.S328996

Hatherley, J. J. (2020). Limits of trust in medical AI. *Journal of Medical Ethics, 46*(7), 478–481. https://doi.org/10.1136/medethics-2019-105935

Håvik, R., Wake, J. D., Flobak, E., Lundervold, A., & Guribye, F. (2019). A conversational interface for self-screening for ADHD in adults. In S. S. Bodrunova, O. Koltsova, A. Følstad, H. Halpin, P. Kolozaridi, L. Yuldashev, A. Smoliarova, & H. Niedermayer (Eds.), *Internet science* (pp. 133–144). Springer International Publishing. https://doi.org/10.1007/978-3-030-17705-8_12.

Hayes, A. M., Beck, J. G., & Yasinski, C. (2012). A cognitive behavioral perspective on corrective experiences. In *Transformation in psychotherapy: Corrective experiences across cognitive behavioral, humanistic, and psychodynamic approaches* (pp. 69–83). American Psychological Association. https://doi.org/10.1037/13747-005.

He, Y., Yang, L., Qian, C., Li, T., Su, Z., Zhang, Q., & Hou, X. (2023). Conversational agent interventions for mental health problems: Systematic review and meta-analysis of randomized controlled trials. *Journal of Medical Internet Research, 25*(1), Article e43862. https://doi.org/10.2196/43862

Heinz, M. V., Mackin, D. M., Trudeau, B. M., Bhattacharya, S., Wang, Y., Banta, H. A., Jewett, A. D., Salzhauer, A. J., Griffin, T. Z., & Jacobson, N. C. (2025). Randomized trial of a generative AI chatbot for mental health treatment. *NEJM AI, 2*(4). https://doi.org/10.1056/aioa2400802

Henson, P., Peck, P., & Torous, J. (2019). Considering the therapeutic alliance in digital mental health interventions. *Harvard Review of Psychiatry, 27*(4), 268–273. https://doi.org/10.1097/HRP.0000000000000224

Herbener, A. B., Klincewicz, M., & Damholdt, M. F. (2024). A narrative review of the active ingredients in psychotherapy delivered by conversational agents. *Computers in Human Behavior Reports, 14*, Article 100401. https://doi.org/10.1016/j.chbr.2024.100401

Ho, F. Y.-Y., Yeung, W.-F., Ng, T. H.-Y., & Chan, C. S. (2016). The efficacy and cost-effectiveness of stepped care prevention and treatment for depressive And/Or anxiety disorders: A systematic review and meta-analysis. *Scientific Reports, 6*(1), Article 1. https://doi.org/10.1038/srep29281

Howe, P. D. L., Fay, N., Saletta, M., & Hovy, E. (2023). ChatGPT's advice is perceived as better than that of professional advice columnists. *Frontiers in Psychology, 14*. https://doi.org/10.3389/fpsyg.2023.1281255 [Brief Research Report].

Howgego, I. M., Yellowlees, P., Owen, C., Meldrum, L., & Dark, F. (2003). The therapeutic alliance: The key to effective patient outcome? A descriptive review of the evidence in community mental health case management. *Australian and New Zealand Journal of Psychiatry, 37*(2), 169–183. https://doi.org/10.1046/j.1440-1614.2003.01131.x

Huibers, M. J. H., Lorenzo-Luaces, L., Cuijpers, P., & Kazantzis, N. (2021). On the road to personalized psychotherapy: A research agenda based on cognitive behavior therapy for depression [hypothesis and theory]. *Frontiers in Psychiatry, 11*. https://doi.org/10.3389/fpsyt.2020.607508

Istepanian, R., Laxminarayan, S., & Pattichis, C. S. (2007). *M-Health: Emerging Mobile health systems*. Berlin: Springer Science & Business Media.

Jacobs, N. (2020). Capability sensitive design for health and wellbeing technologies. *Science and Engineering Ethics, 26*(6), 3363–3391.

Javaid, M., Haleem, A., & Singh, R. P. (2023). ChatGPT for healthcare services: An emerging stage for an innovative perspective. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 3*(1), Article 100105. https://doi.org/10.1016/j.tbench.2023.100105

Jobes, D. A., & O'Connor, S. S. (2009). The duty to protect suicidal clients: Ethical, legal, and professional considerations. In J. L. WerthJr, E. R. Welfel, & G. A. H. Benjamin (Eds.), *The duty to protect: Ethical, legal, and professional considerations for mental health professionals* (pp. 163–180). American Psychological Association. https://doi.org/10.1037/11866-011.

Kang, S.-H., & Gratch, J. (2010). Virtual humans elicit socially anxious interactants' verbal self-disclosure. *Computer Animation and Virtual Worlds*. https://doi.org/10.1002/cav.345. n/a-n/a.

Kang, E., & Kang, Y. A. (2023). Counseling chatbot design: The effect of anthropomorphic chatbot characteristics on user self-disclosure and companionship. *International Journal of Human-Computer Interaction*. https://doi.org/10.1080/10447318.2022.2163775

Karcher, N. R., & Presser, N. R. (2018). Ethical and legal issues addressing the use of Mobile health (mHealth) as an adjunct to psychotherapy. *Ethics & Behavior, 28*(1), 1–22. https://doi.org/10.1080/10508422.2016.1229187

Kaser-Boyd, N. (2015). Threat assessment in homicide/suicide: The duty to warn 1. In C. de Ruiter, & N. Kaser-Boyd (Eds.), *Forensic psychological assessment in practice: Case studies* (1st ed.). Routledge. https://doi.org/10.4324/9781315723310.

Kay, M., Santos, J., & Takane, M. (2011). mHealth: New Horizons for health through Mobile technologies. *World Health Organization, 64*, 66–71.

Kazdin, A. (2018). Expanding mental health services through novel models of intervention delivery. *Journal of Child Psychology and Psychiatry, 60*. https://doi.org/10.1111/jcpp.12937

Kim, T. W., Jiang, L., Duhachek, A., Lee, H., & Garvey, A. (2022). Do you mind if I ask you a personal question? How AI service agents alter consumer self-disclosure. *Journal of Service Research, 25*(4), 649–666. https://doi.org/10.1177/10946705221120232

Klincewicz, M. (2023). Institutional trust in medicine in the age of artificial intelligence. In *The moral psychology of trust (moral psychology of the emotions)*. Rowman & Littlefield.

Klincewicz, M., & Frank, L. (2019). Consequences of unexplainable machine learning for the notions of a trusted doctor and patient autonomy. *CEUR Workshop Proceedings, 2681*.

Knaak, S., Mantler, E., & Szeto, A. (2017). Mental illness-related stigma in healthcare. *Healthcare Management Forum, 30*(2), 111–116. https://doi.org/10.1177/0840470416679413

Kuhail, M. A., Alturki, N., Thomas, J., & Alkhalifa, A. (2024). Human vs. AI counseling: College students' perspectives. *Computers in Human Behavior Reports, 16*, Article 100534. https://doi.org/10.1016/j.chbr.2024.100534

Kuhail, M. A., Alturki, N., Thomas, J., Alkhalifa, A. K., & Alshardan, A. (2024). Human-human vs Human-AI therapy: An empirical study. *International Journal of Human-Computer Interaction*, 1–12. https://doi.org/10.1080/10447318.2024.2385001

Liégeois, A., & Eneman, M. (2012). Ethical aspects of the prevention of suicide in psychiatry. *Ethical Human Psychology and Psychiatry, 14*, 140–149. https://doi.org/10.1891/1559-4343.14.2.140

Lipson, S. K., Zhou, S., Abelson, S., Heinze, J., Jirsa, M., Morigney, J., Patterson, A., Singh, M., & Eisenberg, D. (2022). Trends in college student mental health and help-seeking by race/ethnicity: Findings from the national healthy minds study, 2013–2021. *Journal of Affective Disorders, 306*, 138–147. https://doi.org/10.1016/j.jad.2022.03.038

Loosman, I. (2024). *Rethinking informed consent in mHealth*. Eindhoven University of Technology [Phd Thesis 1 (Research TU/e/Graduation TU/e), Industrial Engineering and Innovation Sciences].

Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology, 5*(1), 96–116. https://doi.org/10.1002/lio2.354

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

Lundh, L.-G. (2014). The search for common factors in psychotherapy: Two theoretical models with different empirical implications. *Psychology and Behavioral Sciences, 3*, 131–150. https://doi.org/10.11648/j.pbs.20140305.11

Mackenzie, C. (Ed.). (2014). *Vulnerability: New essays in ethics and feminist philosophy*. Oxford University Press.

Martinengo, L., Van Galen, L., Lum, E., Nguyen, Q. J., Car, J., & Majeed, A. (2019). Suicide prevention and depression apps' suicide risk assessment and management: A systematic assessment of adherence to clinical guidelines. *BMC Medicine, 17*, 231. https://doi.org/10.1186/s12916-019-1461-z

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

Moitra, M., Santomauro, D., Collins, P. Y., Vos, T., Whiteford, H., Saxena, S., & Ferrari, A. J. (2022). The global gap in treatment coverage for major depressive disorder in 84 countries from 2000–2019: A systematic review and Bayesian meta-regression analysis. *PLoS Medicine, 19*(2), Article e1003901. https://doi.org/10.1371/journal.pmed.1003901

Molli, V. L. P. (2022). Effectiveness of AI-Based chatbots in mental health support: A systematic review. *Journal of Healthcare AI and ML, 9*(9), 1–11.

Murphy, K., Di Ruggiero, E., Upshur, R., Willison, D. J., Malhotra, N., Cai, J. C., Malhotra, N., Lui, V., & Gibson, J. (2021). Artificial intelligence for good health: A scoping review of the ethics literature. *BMC Medical Ethics, 22*(1), 14. https://doi.org/10.1186/s12910-021-00577-8

Nickel, P., Loosman, I., Frank, L., & Vinnikova, A. (2023). Justice and empowerment through digital health: Ethical challenges and opportunities. *Digital Society, 2*. https://doi.org/10.1007/s44206-023-00075-x

Onnela, J.-P., & Rauch, S. L. (2016). Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology, 41*(7), 1691–1696. https://doi.org/10.1038/npp.2016.7

Oosterlaken, I. (2009). *Design for development: A capability approach*.

Ovsyannikova, D., Oldemburgo de Mello, V., & Inzlicht, M. (2025). Third-party evaluators perceive AI as more compassionate than expert humans. *Communications psychology, 3*, 4. https://doi.org/10.1038/s44271-024-00182-6

Pham, K. T., Nabizadeh, A., & Selek, S. (2022). Artificial intelligence and chatbots in psychiatry. *Psychiatric Quarterly, 93*(1), 249–253. https://doi.org/10.1007/s11126-022-09973-8

Philip, P., Micoulaud-Franchi, J.-A., Sagaspe, P., Sevin, E. D., Olive, J., Bioulac, S., & Sauteraud, A. (2017). Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders. *Scientific Reports, 7*(1), Article 42656. https://doi.org/10.1038/srep42656

Pickard, M. D., & Roster, C. A. (2020). Using computer automated systems to conduct personal interviews: Does the mere presence of a human face inhibit disclosure? *Computers in Human Behavior, 105*, Article 106197. https://doi.org/10.1016/j.chb.2019.106197

Pickard, M. D., Roster, C. A., & Chen, Y. (2016). Revealing sensitive information in personal interviews: Is self-disclosure easier with humans or avatars and under what conditions? *Computers in Human Behavior, 65*, 23–30. https://doi.org/10.1016/j.chb.2016.08.004

Rhodes, R. (2001). Understanding the trusted doctor and constructing a theory of bioethics. *Theoretical Medicine and Bioethics, 22*(6), 493–504. https://doi.org/10.1023/A:1014430208720

Rhodes, R. (2020). *The trusted doctor: Medical ethics and professionalism*. New York, NY: Oxford University Press.

Roberts, T., Miguel Esponda, G., Torre, C., Pillai, P., Cohen, A., & Burgess, R. A. (2022). Reconceptualising the treatment gap for common mental disorders: A fork in the road for global mental health? *The British Journal of Psychiatry, 221*(3), 553–557. https://doi.org/10.1192/bjp.2021.221

Rogers, W., Mackenzie, C., & Dodds, S. (2012). Why bioethics needs a concept of vulnerability. *IJFAB: International Journal of Feminist Approaches to Bioethics, 5*(2), 11–38.

Rohrbach, P. J., Dingemans, A. E., Evers, C., Furth, E. F. V., Spinhoven, P., Aardoom, J. J., Lähde, I., Clemens, F. C., Marle, M. E. V., & den, A.-V. (2023). Cost-effectiveness of internet interventions compared with treatment as usual for people with mental disorders: Systematic review and meta-analysis of randomized controlled trials. *Journal of Medical Internet Research, 25*(1), Article e38204. https://doi.org/10.2196/38204

Rüsch, N., Angermeyer, M. C., & Corrigan, P. W. (2005). Mental illness stigma: Concepts, consequences, and initiatives to reduce stigma. *European Psychiatry, 20*(8), 529–539. https://doi.org/10.1016/j.eurpsy.2005.04.004

Sah, Y. J., & Peng, W. (2015). Effects of visual and linguistic anthropomorphic cues on social perception, self-awareness, and information disclosure in a health website. *Computers in Human Behavior, 45*, 392–401. https://doi.org/10.1016/j.chb.2014.12.055

Santoni De Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology, 34*(4), 1057–1084. https://doi.org/10.1007/s13347-021-00450-x

Schmietow, B., & Marckmann, G. (2019). Mobile health ethics and the expanding role of autonomy. *Medicine, Healthcare & Philosophy*, 623–630. https://doi.org/10.1007/s11019-019-09900-y

Schöbel, S., Schmitt, A., Benner, D., Saqr, M., Janson, A., & Leimeister, J. M. (2023). Charting the evolution and future of conversational agents: A research agenda along five waves and new frontiers. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-023-10375-9

Sedlakova, J., & Trachsel, M. (2022). Conversational artificial intelligence in psychotherapy: A new therapeutic tool or agent? *The American Journal of Bioethics*, 1–10. https://doi.org/10.1080/15265161.2022.2048739

Shahriar, S., & Hayawi, K. (2023). *Let's have a chat! A conversation with ChatGPT: Technology, applications, and limitations*. https://doi.org/10.47852/bonviewAIA3202939

Siddals, S., Torous, J., & Coxon, A. (2024). "it happened to be the perfect thing": Experiences of generative AI chatbots for mental health. *Npj Mental Health Research, 3*(1), 48.

Simon, R. (2002). Suicide risk assessment: What is the standard of care? *The journal of the American Academy of Psychiatry and the Law., 30*, 340–344.

Simonsen, J., & Robertson, T. (Eds.). (2013). *Routledge international handbook of participatory design* (Vol.711). New York: Routledge.

Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2021). My chatbot companion - A study of human-chatbot relationships. *International Journal of Human-Computer Studies, 149*, Article 102601. https://doi.org/10.1016/j.ijhcs.2021.102601

Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2022). A longitudinal study of human–chatbot relationships. *International Journal of Human-Computer Studies, 168*, Article 102903. https://doi.org/10.1016/j.ijhcs.2022.102903

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*, 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B., Sedoc, J., Derubeis, R. J., Willer, R., & Eichstaedt, J. C. (2024). Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation. *Mental Health Research, 3*(1). https://doi.org/10.1038/s44184-024-00056-z

Steyn, C. R. (2006). Breach of confidentiality and the duty to warn in medical law: Examples from clinical psychiatry. *Journal for Juridical Science, 31*, 134–142.

Sweeney, C., Potts, C., Ennis, E., Bond, R., Mulvenna, M. D., O'Neill, S., Malcolm, M., et al. (2021). Can chatbots help support a person's mental health? Perceptions and views from mental healthcare professionals and experts. *ACM Transactions on Computing for Healthcare, 2*(3), 1–15. https://doi.org/10.1145/3453175

Ten Have, H. (2015). Respect for human vulnerability: The emergence of a new principle in bioethics. *Journal of Bioethical Inquiry, 12*, 395–408.

Torous, J., Gershon, A., Hays, R., Onnela, J.-P., & Baker, J. T. (2019). Digital phenotyping for the busy psychiatrist: Clinical implications and relevance. *Psychiatric Annals, 49*(5), 196–201. https://doi.org/10.3928/00485713-20190417-01

Twenge, J. M., Cooper, A. B., Joiner, T. E., Duffy, M. E., & Binau, S. G. (2019). Age, period, and cohort trends in mood disorder indicators and suicide-related outcomes in a nationally representative dataset, 2005–2017. *Journal of Abnormal Psychology, 128*(3), 185–199. https://doi.org/10.1037/abn0000410

van der Storm, S. L., Jansen, M., Meijer, H. A. W., Barsom, E. Z., & Schijven, M. P. (2023). Apps in healthcare and medical research: European legislation and practical tips every healthcare provider should know. *International Journal of Medical Informatics, 177*, Article 105141. https://doi.org/10.1016/j.ijmedinf.2023.105141

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine, 15*(11), Article e1002689.

Vigo, D. V., Kestel, D., Pendakur, K., Thornicroft, G., & Atun, R. (2019). Disease burden and government spending on mental, neurological, and substance use disorders, and self-harm: Cross-sectional, ecological study of health system response in the americas. *The Lancet Public Health, 4*(2), e89–e96. https://doi.org/10.1016/S2468-2667(18)30203-2

Vowels, L. M. (2024). Are chatbots the new relationship experts? Insights from three studies. *Computers in Human Behavior: Artificial Humans, 2*(2), Article 100077. https://doi.org/10.1016/j.chbah.2024.100077

Wahde, M., & Virgolin, M. (2022). Conversational agents: Theory and applications. *Handbook of Computer Learning and Intelligence, 1*. https://doi.org/10.48550/arXiv.2202.03164

Wampold, B. E., & Flückiger, C. (2023). The alliance in mental health care: Conceptualization, evidence, and clinical applications. *World Psychiatry, 22*(1), 25–41. https://doi.org/10.1002/wps.21035

Wentzel, J., Vaart, R. van der, Bohlmeijer, E., & Gemert-Pijnen, J. V. van (2016). Mixing online and face-to-face therapy: How to benefit from blended care in mental health care. *JMIR Mental Health, 3*. https://doi.org/10.2196/mental.4534

Yang, Y., Tavares, J., & Oliveira, T. (2024). A new research model for artificial intelligence–based well-being chatbot engagement: Survey study. *JMIR Hum Factors, 11*, Article e59908. https://doi.org/10.2196/59908