



OPEN ACCESS

EDITED BY

Ward Van Breda,
VU Amsterdam, Netherlands

REVIEWED BY

Eduardo L. Bunge,
Palo Alto University, United States
Warren Mansell,
The University of Manchester, United Kingdom

*CORRESPONDENCE

J. P. Grodniewicz
✉ j.grodniewicz@gmail.com
Mateusz Hohol
✉ mateusz.hohol@uj.edu.pl

RECEIVED 20 March 2023

ACCEPTED 15 May 2023

PUBLISHED 01 June 2023

CITATION

Grodniewicz JP and Hohol M (2023) Waiting for a digital therapist: three challenges on the path to psychotherapy delivered by artificial intelligence.

Front. Psychiatry 14:1190084.

doi: 10.3389/fpsy.2023.1190084

COPYRIGHT

© 2023 Grodniewicz and Hohol. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Waiting for a digital therapist: three challenges on the path to psychotherapy delivered by artificial intelligence

J. P. Grodniewicz¹* and Mateusz Hohol¹*

Copernicus Center for Interdisciplinary Studies, Jagiellonian University, Kraków, Poland

Growing demand for broadly accessible mental health care, together with the rapid development of new technologies, trigger discussions about the feasibility of psychotherapeutic interventions based on interactions with Conversational Artificial Intelligence (CAI). Many authors argue that while currently available CAI can be a useful supplement for human-delivered psychotherapy, it is not yet capable of delivering fully fledged psychotherapy on its own. The goal of this paper is to investigate what are the most important obstacles on our way to developing CAI systems capable of delivering psychotherapy in the future. To this end, we formulate and discuss three challenges central to this quest. Firstly, we might not be able to develop effective AI-based psychotherapy unless we deepen our understanding of what makes human-delivered psychotherapy effective. Secondly, assuming that it requires building a therapeutic relationship, it is not clear whether psychotherapy can be delivered by non-human agents. Thirdly, conducting psychotherapy might be a problem too complicated for narrow AI, i.e., AI proficient in dealing with only relatively simple and well-delineated tasks. If this is the case, we should not expect CAI to be capable of delivering fully-fledged psychotherapy until the so-called “general” or “human-like” AI is developed. While we believe that all these challenges can ultimately be overcome, we think that being mindful of them is crucial to ensure well-balanced and steady progress on our path to AI-based psychotherapy.

KEYWORDS

artificial intelligence, mental health chatbots, psychotherapy, therapeutic alliance, narrow vs. general AI, large language models, Cognitive Behavioral Therapy (CBT), conversational agents

1. Introduction

The mental health crisis is arguably among the most important global challenges that we are currently facing (1–3). Responding to it would require developing large-scale, high-quality, accessible, and affordable mental health care solutions (4). We might not be able to do this without benefitting from cutting-edge technology, including Artificial Intelligence (AI).

In recent years, AI has begun to be implemented in multiple domains of mental health care (5–7). AI-based solutions are used to improve the diagnosis of depression (8–10) and schizophrenia (11–13) and predicting treatment outcomes (14–17). Intelligent robots work with children with autism spectrum disorders (18) and elderly people suffering from dementia (19). Virtual reality avatars help patients confront their auditory hallucinations (20, 21). The list goes on.

One subdomain of mental health care where implementation of AI technology is both particularly challenging and promising is psychotherapy or “talk therapy” (22–31). It is difficult to find an uncontroversial definition of talk therapy acceptable to representatives of all therapeutic traditions. Nevertheless, as a first approximation, we can appeal to the following characterization offered by American Psychological Association (APA) (32), according to which psychotherapy is:

communication between patients and therapists that is intended to help people: (i) find relief from emotional distress, as in becoming less anxious, fearful or depressed, (ii) seek solutions to problems in their lives, such as dealing with disappointment, grief, family issues, and job or career dissatisfaction, and (iii) modify ways of thinking and acting that are preventing them from working productively and enjoying personal relationships.¹

The APA characterization goes on to differentiate therapy from “talking with a friend.” While both a therapist and a friend may be willing to listen about our problems, only therapists are “trained professionals with specialized education and experience in understanding psychological problems” (32). Moreover, in contrast with friendship, therapy is non-symmetrical, and focuses solely on the client’s well-being. Finally, therapy takes place in a structured setting, typically there is an agreement between the client and therapist regarding regular meeting times, the length of each meeting, etc.

Given the rapid development of ever more impressive AI technologies, it is only natural to wonder whether AI already is, or will be in the future, able to conduct psychotherapy, and thus challenge the traditional conceptualization of psychotherapy as the relationship between two flesh-and-blood persons in which one cures the other. The closest we get to AI-based psychotherapy these days are the interventions delivered by mental health chatbots, many of which are based on relatively simple dialogue systems (34). In the next section (Section 2), we briefly discuss why several authors suggest that – even though they are sometimes oversold as offering fully fledged psychotherapy – currently available chatbots are falling short of achieving this goal. But what about chatbots available in 2, 5, or 10 years? As proved by the impressive performance of ChatGPT produced by OpenAI-Microsoft, interactions with CAI based on so-called “Large Language Models” (LLMs) start becoming deceptively similar to conversations with another human being. Moreover, while some authors are cautious about the abilities of CAI based on LLMs (35), others go as far as to claim that appropriately prompted CAIs are able to perform complex reasoning (36) or even manifest abilities similar to what psychologists call Theory of Mind, i.e., the ability to assign mental states (such as beliefs, desires, and intentions) to other agents (37, however, see 38). Does this mean that fully fledged psychotherapy delivered by CAI is around the corner? This is not so

simple. On our way to developing CAI systems capable of delivering psychotherapy, we will encounter problems and obstacles that might be impossible to overcome by simply increasing the computational power of AI algorithms or training data sets of LLMs. The goal of this paper is to characterize three of them.

The first (discussed in Section 3) is *The Problem of a Confused Therapist*. The gist of the problem is that different therapeutic traditions conceptualize the process of psychotherapy differently, often disagreeing about what a therapist should do while conducting therapy. Moreover, there is heated discussion regarding the effectiveness of different therapeutic approaches, and it is unclear whether we will be able to develop an artificial system delivering the processes, techniques, and interventions that make psychotherapy effective unless we first better understand which processes, techniques, and interventions make psychotherapy effective. Clinical psychologists have tried to answer this question for the last 100 years, and we still seem to be far away from a fully satisfactory solution. Part of the problem is that it is not obvious to what extent the active ingredients in effective psychotherapy are specific techniques and interventions as opposed to such common factors as, e.g., a supportive therapeutic relationship.

This leads to the second problem – *The Problem of a Non-human Therapist* (Section 4). It has been traditionally assumed that psychotherapeutic care takes place within the frame of a relationship between two (or more) people. Can the “human aspect” of psychotherapy be reconstructed in the relationship between a person and an artificial system, e.g., by breaking it down into constituents such as trust, empathy, respect, mutual understanding, etc., which could later be emulated by an AI?

The third problem (Section 5) is *The Problem of a Narrowly Intelligent Therapist*. Despite the hype and enthusiasm around AI research, at least in the proximate future, we are stuck with narrow AI technologies. Narrow (as opposed to general) AI systems are very effective but only in performing well-delineated and clearly constrained tasks, such as image recognition, speech processing and generation, or even driving a car. The question is whether conducting psychotherapy is a task that can be so clearly delineated (or whether it can be divided into a set of such tasks). If not, it does not seem that we will be able to develop AI systems conducting fully fledged psychotherapy unless we first build a general AI, i.e., AI which – very much like a human – would be able to apply its intelligence to an unrestricted set of tasks.

2. Existing mental health chatbots

With more than 10,000 available options, the market of mental health apps is booming (39). Most of them implement relatively simple functionalities, such as mood, progress, and medication tracking, journaling, or easy access to prerecorded guided meditations and breathing exercises. However, in this section, we want to focus on the subset of most pioneering among these apps – mental health chatbots (6, 24, 26–31, 40).

Mental health chatbots are dialogue systems designed to interact with a human in real time in a way that mimics human–human conversation. To get a better grip on how mental health chatbots work, it will be useful to take a closer look at a fragment of an exemplary conversation with Woebot – one of the most popular and advanced

1 Another relatively universal definition of psychotherapy comes from Norcross (33): “Psychotherapy is the informed and intentional application of clinical methods and interpersonal stances derived from established psychological principles for the purpose of assisting people to modify their behaviors, cognitions, emotions, and/or other personal characteristics in directions that the participants deem desirable” [p. 218].

chatbots designed to implement principles of Cognitive Behavioral Therapy (CBT) (41).²

- [1] Woebot: Can you say more about what's going on?
 [2] User: I'm feeling bad.
 [3] Woebot: This sounds like a problem with feeling low, does that sound right to you?
 [4] User: [picking one of the available continuations] Yeah, that's right.
 [5] Woebot: Sadness can make the smallest tasks seem like unclimbable mountains, but I want you to know: I'm here to help and we can get through this together.
 [6] User: [picking the only available continuation] Yeah.
 [7] Woebot: There's two ways we can go about this. I can help you challenge your thinking, or we can focus on things you can do right now.
 [8] User: [picking one of the available continuations] Thought challenger.
 [9] Woebot: When you are feeling low, it can be helpful to understand how your thoughts affect your emotions. So we are going to take a close look at your thinking to help you feel better.
 [10] User: [picking the only available continuation] Okay.
 [11] Woebot: Before we dive in, a quick reminder: your mind is a powerful, incredible thing! It can solve problems, make plans, and process information at warp speed. But sometimes, our turbo-charged brains jump to conclusions that aren't 100% true.
 [12] User: [picking one of the available continuations] Makes sense.
 [13] Woebot: Psychologists call these automatic negative thoughts – harsh ideas or statements that pop into your mind as you navigate daily life. Some people compare them to put-downs from a tiny character, like a villain or bully, that lives inside their head.
 [14] User: [picking the only available continuation] I see.
 [15] Woebot: Often these thoughts can be distorted, which means they make things feel worse or more extreme than they actually are. But once you learn to catch these thoughts and check them for distortions, you can change them so they are fairer, more balanced, and truer to reality.

Already in this short conversation, Woebot mimics a lot of elements of an actual therapy session. Just as a human therapist might do, it makes sure that it has understood the user correctly [3], normalizes the user's experience ("Sadness can make the smallest tasks seem like unclimbable mountains), and suggests a form of the therapeutic alliance ("we can get through this together") [5]. It also offers elements of psychoeducation by explaining that thoughts influence our emotions [9] and that our brains fall prey to cognitive errors [11], and by outlining the concept of "automatic negative thoughts" with the metaphor of "put-downs" from a "villain or bully" [13]. Finally, it introduces the basic idea behind one of the main therapeutic processes of CBT, i.e., cognitive restructuring (42), by suggesting that we should try to "catch," "check," and "change" our

distorted thoughts [15]. In the further course of the conversation, Woebot guides the user through the process of restructuring, inviting them to recognize and revise distortions underlying a particular thought. For example, when a user inputs "Nobody loves me," the exemplary conversation continues:

- [16] Woebot: Consider your thought "Nobody loves me." Does it assume that because one bad thing happened, things are going to stay bad from here on out?
 [17] User: [picking one of the available continuations] Yes.
 [18] Woebot: Good catch. This distortion is called overgeneralization. If you make assumptions about your entire life based on a single event or situation, you are probably overgeneralizing.

Finally, Woebot invites the user to reformulate the thought in a way omitting these distortions, thus concluding a short intervention implementing the process of cognitive restructuring.

Several ethical problems involved in the use of mental health chatbots have been already raised (28–30, 43–46). They include, e.g., (i) concerns regarding data privacy, (ii) the risk of bypassing rather than fighting the stigma related to mental health issues by encouraging users to keep their struggle in the privacy of the interaction with their phone, and (iii) the lack of control over feedback and recommendations users receive from the app. These and other related risks are already discussed in considerable detail (28–30, 43–46) and we will not repeat these arguments here. Instead, we will focus on the question central to our current discussion: "what exactly do mental health chatbots do?"

In review articles devoted to therapeutic CAI we read that among the functions of mental health chatbots there is "delivering evidence-based psychological interventions" (26) [p. 1]; "providing cognitive behavioral therapy (CBT)" (31) [p. 459] and that "[t]he most common use of chatbots was delivery of therapy, training, and screening" (40) [p. 6]. Companies developing such chatbots are increasingly cautious not to characterize the services offered by chatbots as psychotherapy. For example, on its website, Woebot is presented as "your personal mental health ally," not a therapist (41). However, the same website mentions that it helps "deliver individual support through interactive and easy-to-use therapeutic solutions" and that "There's no such thing as appointments or waiting rooms here," which clearly indicates a visit to one's therapist as opposed to, e.g., using a self-help book, as the appropriate comparison class for Woebot. In a similar fashion, another popular product of this type – Wysa – is presented as "an AI chatbot that leverages evidence-based cognitive-behavioral techniques (CBT) to make you feel heard" (47). It is unclear whether "leveraging" therapeutic techniques equals "using" them and whether using therapeutic techniques is assumed to equal delivering psychotherapy. Among the benefits of using Tess – yet another mental health chatbot developed by X2 Foundation – we find "Effective: Most people report that they prefer Chat With Tess over traditional therapy..." and "Affordable: Support from Tess is 98% cheaper than face-to-face therapy" (48). Again, without explicitly calling Tess a "therapist," X2 Foundation clearly suggests that it is appropriate to compare using their chatbot to attending human-delivered psychotherapy.

The lack of transparency regarding the actual nature of services provided by chatbots is probably most visible in what we may call "The Efficacy Overflow Argument," often implicitly conveyed in the marketing of mental health chatbots (29). Here is the general form of the argument:

² Based on Version 4.8.1.(214) of Woebot app. Conversational turns are numbered for reference in square brackets (so that they were not to be confused with literature references provided in parentheses).

The efficacy overflow argument: (1) Chatbot C is based on the principles of a psychotherapeutic approach P. (2) There is evidence for efficacy of P. (3) Therefore, we should consider C to be evidence-based.

The argument is invalid. The fact that a given psychotherapeutic approach is effective when the therapy is conducted by a well-trained and experienced therapist over the course of multiple one-to-one sessions does not mean that the techniques or interventions based on the principles of this approach will be effective psychotherapy when administered by a chatbot.

At the same time, a growing body of research shows that interactions with chatbots can contribute to improving their users' mental health and quality of life, especially if the specific needs of the user match with chatbot capabilities (22, 24, 27, 31, 34, 40, 49–56). While these studies differ in terms of evaluation methods, in a recent meta-analysis of 32 randomized controlled trials, Yuhao He and colleagues (57) found that CAIs proved to be effective, in particular on a proximal time scale, in reducing depressive and anxiety (both general and specific) symptoms, preventing stress, general distress, and negative affect, and improving well-being. In light of these findings, the authors conclude that, “in the post-epidemic and digital eras, CAIs will likely play a significant role and contribute significantly to the new health transformation, in our care” [p. 15].

But can we say that the kind of psychological support offered by chatbots is already equivalent to psychotherapy, considered traditionally as a relationship between two persons or agents in which one of them heals or cures the other one (30, 45)? In a recent paper, Jana Sedlakova and Manuel Trachsel (28) argue that we should think about a conversational chatbot as “new artifacts” lying on the spectrum between therapeutic tools and therapists. According to these authors, chatbot is not a therapist because it is not a subject or agent. But it is not a mere tool either because it “might be experienced and treated as if it was a subject or agent” (28) [p. 4]. Unlike human therapists, existing chatbots cannot engage in normal human discursive practices, characterized by the ability of giving and asking for reasons, as well as understanding and explaining the concepts one uses. As such, according to Sedlakova and Trachsel, they cannot facilitate the acquisition of insight or self-understanding – one of the central elements of a psychotherapeutic process. On the other hand, as we argue in (58), interactions with a chatbot implementing CBT techniques might put a user in a better position to recognize “the connections between one's emotions, motivations, thoughts, and behavior, past and present, including one's interpretations of and relations with others” (59) (pp. 154–5), which is arguably at least an element of the process of acquiring self-understanding.

To sum up, we think that Sedlakova and Trachsel and other authors are right in raising the question of whether existing chatbots deliver support equivalent to traditional psychotherapy. We also agree with them that, in the foreseeable future, chatbots should be used cautiously; ideally as supplements for human delivered psychotherapeutic care [cf. (5, 26, 31)]. On the other hand, having in mind the growing effectiveness of existing mental health chatbots in reducing psychiatric symptoms and improving the well-being of their users as well as witnessing the current boom of CAI technologies along with their expected development over the next few years, it is worth considering what conditions would have to be met and what challenges overcome for us to be able to call an interaction with an

artificial intelligence “psychotherapy.” This is the goal for the rest of this paper.

3. The problem of a confused therapist

The field of psychotherapeutic care is by no means monolithic. Prochaska and Norcross (60) estimate that there are now more than 500 different psychotherapeutic approaches. Most of these approaches belong to one of the main psychotherapeutic traditions, e.g., psychoanalytic/psychodynamic, existential, person-centered, behavioral, cognitive, etc. Principles underlying some of these traditions are more or less compatible with each other, which enabled the creation of therapeutic modalities benefitting from more than one of them. A prominent example here is Cognitive Behavioral Therapy (CBT) – undoubtedly one of the most popular modern therapeutic orientations (61). Principles underlying some other traditions (e.g., the psychoanalytic and the behavioral tradition) seem so far apart that it is much more difficult to think about their productive combination [which is not to say that such attempts have not been made (62)].

Different therapeutic traditions conceptualize human mental suffering differently. For example, while psychoanalytic tradition focuses on unresolved internal conflicts, cognitive tradition focuses on maladaptive beliefs, and patterns of thinking. Such theoretical differences result, in turn, in different repertoires of clinical processes and techniques – to put it simply, therapists working in different traditions do different things. For example, while we may expect a psychoanalyst to work with their client using free associations, dream interpretation, or analysis of transference, a CBT therapist will rather appeal to such tools as cognitive restructuring or exposure. In sum, specific therapeutic orientation influences the therapist's way of “(a) generating hypotheses about a client's experience and behavior, (b) formulating a rationale for specific treatment interventions, and (c) evaluating the ongoing therapeutic process” (63) [p. 412].

On top of that, the question that has loomed over the whole field for the last 100 years is whether different psychotherapeutic orientations (to the extent to which they are effective) are effective thanks to their specific features, or the so called “common” or “non-specific” factors. Since the publication of Rosenzweig (64), different authors suggested dozens of lists of common factors potentially responsible for making different therapies effective independently of what specific techniques they involve (65–70). For example, Weinberg (71) discusses such common factors as: (i) the therapeutic relationship between a therapist and a client/patient, (ii) expectations of therapeutic success, (iii) client's confronting or facing the problem they struggle with, (iv) experience of mastery or control over the problematic issue, and (v) attribution of therapeutic success or failure to internal causes (e.g., changes in client's coping skills) rather than external causes (e.g., therapist's abilities and techniques).

This leads to the first problem on our way towards a fully fledged AI-based psychotherapy:

The Problem of a Confused Therapist: Can we develop artificial systems capable of conducting effective psychotherapy, given our limited understanding of the necessary components of a therapeutic process and factors that make psychotherapy effective?

This problem is relevant to the quest of achieving fully fledged AI-based psychotherapy but not specific to it. We face the same problem in the case of training of future therapists in general. Which orientation should they choose? Should they become Cognitive Behavioral, Gestalt, or psychodynamic therapists? Or should they pick yet something else from the plethora of options?

Research suggests that the decision to choose one's therapeutic orientation is based, among others, on such factors as personality (72), individual learning styles (73), and value preferences (74). But in the case of developing artificial therapists, we will have to make this decision for them, either by hard-wiring them with principles of a given therapeutic modality or by training AI algorithms on data from sessions in which a particular therapeutic modality is implemented (75). Moreover, assuming that we would like to train our algorithms on data coming from successful psychotherapies, we will have to decide how we want to understand "success" in psychotherapy. Is it identical with symptom reduction? If yes, which ones and measured on what scale? Or maybe we should rather identify it with the improvement of a client's functioning and general well-being (76)?

Furthermore, each therapist faces the question of whether they should be faithfully applying the methods of a therapeutic orientation they have been trained in or whether they should mix it with other methods, techniques, and processes that they find fit. According to Prochaska and Norcross (60), most therapists declare that they integrate methods from different therapeutic orientations. Should artificial therapists also be integrative or eclectic? Should we thus – on principle – train them on data from different therapeutic traditions? Additionally, each therapist must individually find their balance between specific therapeutic techniques and methods they apply during session, and less specific work such as nurturing clients' sense of mastery and control. How should we go about achieving this balance in the case of artificial therapists? Prochaska and Norcross (60) point out that "Without a guiding theory or system of psychotherapy, clinicians would be vulnerable, directionless creatures bombarded with literally hundreds of impressions and pieces of information in a single session" [p. 4]. This is equally true about artificial therapists.

On the other hand, integrative approaches, "characterized by dissatisfaction with single-school approaches and a concomitant desire to look across school boundaries to see what can be learned from other ways of conducting psychotherapy" in order "to enhance the efficacy, efficiency, and applicability of psychotherapy" (77) [p. 4], may prove useful in our future attempts to develop new, and improve existing, therapeutic interventions delivered by CAIs. A notable example of one such attempt is the chatbot MYLO (78), which implements the core principles of a transdiagnostic, integrative therapy called Method of Levels (MOL) (79).³ MOL is based on an all-encompassing psychological theory: Perceptual Control Theory (80, 81), according to which the most important principle guiding life – from the level of basic biological functioning all the way up to mental health and well-being – is control. Psychological distress experienced by people seeking psychotherapeutic help results from the emergence of an internal conflict, which triggers a loss of control. According to the assumptions underlying MOL, all that clients need

to be in a position to resolve such a conflict and regain control is for someone to "(1) help them to talk about the problem at length, in detail and in the present moment, thereby sustaining their attention to it, and (2) to notice disruptions in their speech and behaviour as they describe the problem, such that the client can shift their attention to aspects of the problem they may otherwise have missed" (82) [p. 140]. At least to some extent, this intervention has been implemented in the chatbot MYLO, which simply asks its users a series of questions, thereby creating a context in which a client can explore and resolve their internal conflict. Chatbots such as MYLO, if effective, can constitute proof of concept that solving The Problem of a Confused Therapist can proceed not by creating more sophisticated dialogue systems capable of deliver complicated therapeutic procedures but by relying on relatively simple therapeutic interventions. Even though the preliminary results are promising (78, 83, 84), more comprehensive studies on bigger and more diversified groups of users are in order.

Simultaneously, judging by the state of the field of mental health chatbots, the therapeutic orientation considered to be most promising for AI-based therapy is CBT. CBT, as delivered by a well-trained human specialist, is effective in the treatment of a broad set of diagnoses (85). It is also among the forms of therapy that bring positive results in a relatively short time. Moreover, the basic idea of CBT – that "self-relevant thoughts, evaluations, and beliefs are key contributors to the development and persistence of psychopathological states" (42) [p. 23] – is simple and elegant. At least to some extent, CBT can be broken down into a set of techniques that focus on identifying, challenging, and substituting such maladaptive thoughts and beliefs. This makes CBT relatively simple to operationalize (86). Thus, the assumption of many chatbot developers is that as long as chatbots help their users identify, challenge, and substitute their maladaptive cognitions – together with some additional skills and mindfulness training, and behavior activation – they *deliver* CBT (87).

But – many psychotherapists would suggest – even in the case of CBT, therapy is more than just techniques (88, 89). In fact, a whole chapter of one of the founding CBT textbooks (90) is devoted to the therapeutic relationship. Such a relationship is characterized by "warmth, accurate empathy, and genuineness" (90) [p. 45]. This leads us to the second problem on the way towards fully fledged AI-based psychotherapy.

4. The problem of a non-human therapist

As noted in Section 2, psychotherapy has traditionally been framed as a relationship between two persons or agents. These two characteristics are not on par. While we might be inclined to reserve the term "person" for human persons/people (91) it is less controversial to speak about "artificial agents," assuming that "an agent is a being with the capacity to act, and 'agency' denotes the exercise or manifestation of this capacity" [(92), cf. (93)]. Existing mental health chatbots are obviously not persons, and probably not even agents, given how restricted and limited their capacity to act. But this might not be the case for the artificial therapists of the future. Therefore, the second challenge on the way towards AI-based psychotherapy is the following:

³ We are grateful to a reviewer for drawing our attention to MYLO and MOL.

The Problem of a Non-human Therapist: Can a non-human agent conduct psychotherapy, given that, additionally (or primarily) to delivering specific techniques, it requires building a therapeutic relationship?

According to most authors, psychotherapy is much more than the delivery of specific techniques. It is first and foremost an interpersonal relationship (30, 45, 94, 95). This raises concerns regarding the possibility of fully fledged AI-based psychotherapy (26, 28–30, 45, 96).

However, instead of thinking about the lack of human–human relationship as an insurmountable obstacle, it might be better to think about it as a challenge. Ideally, everyone struggling with a mental health problem would have easy access to a well-trained human therapist, who will be additionally able to devote their clients as much time as necessary. Unfortunately, with a global median of 13 mental health workers per 100,000 population (62.2 in high-income countries) (97), and the estimate of 70% of people with mental illness receiving no treatment from health care staff (98), this scenario is unrealistic. Therefore, without neglecting the importance of human–human relationship, and without abandoning the efforts to increase the number of mental health workers *per capita*, we should actively seek alternative solutions.

It seems that there are three practical strategies to choose from, when confronted with The Problem of Non-human Therapist. Here, we will just put these strategies forward, leaving the task of their careful assessment for another occasion.

The first strategy – *Deflation* – involves deflating the role of a therapeutic relationship in the therapeutic process. Sure – a supporter of this strategy can say – when the psychotherapy involves two people, a deep relationship is likely to appear between them. In fact, this would be true about all prolonged interactions between two people – this is a part of our social set-up. The therapeutic relationship differs from other relationships, such as friendship, and a skilled therapist will steer it in such a way that it is most beneficent to achieving the therapeutic goals. But in the case of artificial therapists, such a relationship might be missing. We used to think about psychotherapy as involving a therapeutic relationship, because we used to think about it as involving humans. But the reality of psychotherapy of the future might be different, and we should not be stuck to our old standards, according to which therapeutic relationship is necessary for the psychotherapeutic process to occur. Maybe, instead of focusing on what is missing, we should focus on what other resources we have? It is true that the contribution of a therapeutic relationship to the overall efficacy of psychotherapy will be lost. But maybe we can regain it in other ways? For example, despite not being able to build a therapeutic relationship, artificial therapists may be much more skilled and consistent in delivering therapeutic techniques than humans, and thus at least as effective as human therapists.

One more thing worth keeping in mind in this context is that we already know some helpful psychological interventions that do not require a therapeutic relationship. A notable example is various writing techniques, which turn out – at least in certain populations – to be effective in reducing psychological distress (99). Therefore, e.g., Carey et al. (100), point out that while a therapeutic relationship might be a key component of effective psychological intervention (they go as far as to suggest that “it might be sufficient” for achieving the positive therapeutic effect; p. 48), it is not necessary. Even if it turns out that it is impossible to build a therapeutic relationship between a user and a

CAI, we still might focus on training chatbots to use interventions that do not require such a relationship.

The second strategy is *Mimicry* [cf. (28)]. According to this strategy, what is important is not whether there is a therapeutic relationship, but whether the client *thinks* that there is such a relationship. In Section 2, we quoted Woebot saying: “I’m here to help and we can get through this together,” even though Woebot is not “together” with us in our mental health struggle any more than we and our umbrella are “together” in the rain. Mimicry seems to be the default strategy in modern chatbot development.

In this context, the focus is often shifted from the thicker concept of the therapeutic relationship to a somewhat more technical notion of *therapeutic alliance*. According to a classical characterization offered by Edward Bordin (101), therapeutic alliance consists of (a) an agreement on therapeutic goals, (b) an assignment of therapeutic tasks, and (c) the development of bonds.⁴ While it might be impossible to achieve a genuine therapeutic relationship (based on warmth, empathy, and acceptance) with a CAI, at least certain aspects of the therapeutic alliance might be reconstructed in an interaction with a chatbot. Thus, following Bordin (101), Kaveladze and Schueller (102) define the digital therapeutic alliance (DTA) as “a user-perceived alliance (composed of a bond, agreement on the tasks directed toward improvement, and agreement on therapeutic goals)” [*emphasis added*, pp. 88–89], which, at least to some extent, can occur in interaction with CAI. To date, at least two psychometric tools have been proposed to conceptualize and measure this aspect of interacting with chatbots (103): Mobile Agnew Relationship Measure (mARM) (104) and Digital Working Alliance Inventory (D-WAI) (105, 106). In particular, the latter – D-WAI – includes items related to therapeutic goals (e.g., “I trust the app to guide me towards my personal goals”), tasks (e.g., “I believe the app tasks will help me to address my problem”), and bonds (e.g., “The app supports me to overcome challenges”), as a way of measuring the strength of therapeutic alliance in the digital context.

In general, there is now an increasing body of research on the digital therapeutic alliance, and its overall influence on the efficacy of help provided by mental health chatbots (51, 52, 54, 56, 102–112). These studies, including randomized controlled trials, reveal varying strengths of the effects: from small up to comparable or even outperforming these found in therapy delivered by humans (as measured on traditional scales designed for assessing the strength of working alliance formed between a client and a human therapist). Thus, we can conclude that existing chatbots allow users to establish DTA, including bonding and agreeing on tasks and goals, which may contribute to reducing symptoms. Considering rapid advances in mental health chatbot development driven by studies in the field of human-computer interaction, we can expect that levels of DTA obtained in questionnaires such as D-WAI or mARM will increase. Most important future improvements will likely include personalization, better adaptation of chatbots to the user’s personality, and better simulation of human characteristics (57, 108, 113–115).

⁴ See also (30), where the alliance is elucidated as “a process where the patient and the therapist work together to determine the goals of treatment based on the patient’s existing problems and expectations from psychotherapy. Thinking together, they identify the steps to achieve that goal, forming a connection in the process” (p. 156).

Nevertheless, it is worth keeping in mind the limitations of such an approach. While psychometric tools such as mARM and D-WAI allow us to compare levels of alliance one achieves with human therapists with those one achieves with a mental health chatbot designed within the same conceptual framework, the current understanding of the DTA phenomenon is still restricted and requires further research with novel measurements. According to Lederman and D'Alfonso:

... given that such measures are more or less based on existing measures of the traditional therapeutic alliance and simply replace “therapist” with “app,” with possibly a few other minor modifications, ultimately such an approach seems unsatisfactory or incomplete, as it does not account for the possibility of certain nuances, particularities, and complexities that could arise in the context of digital interventions. Furthermore, ... one would expect that not all aspects of a traditional therapeutic alliance will necessarily apply to a DTA, and that there may also be dimensions of alliance in the digital context that are not accounted for in traditional therapeutic alliance models (103) [p. 2].

Another problem worth examining is that, while mimicry might increase users' engagement at the beginning, it might also have a detrimental effect when things do not go well, and the users' situation is not improving despite their efforts. In such a case feeling *as if somebody cares*, might not be enough. Moreover, successfully mimicking a therapeutic relationship might be very difficult, given that it involves multiple aspects and degrees “from a sense of being provided for (the therapist will take care of me), to a safe haven (the therapist will protect me), to a solid base (life is predictable here), to a sense of coherence (the therapist understands me), to being attuned to (the therapist and I are one)” (71) [p. 49]. On the other hand, already the famous ELIZA, a simple computer program designed to mimic an interaction with a Rogerian psychotherapist, was supposedly good enough to successfully trick at least some of its users into thinking that they talk to a human therapist (116). Last but not least, mimicry raises ethical problems (28–30). In short, mimicry is a form of deception, and it is more effective the more its receivers are deceived.

The last strategy – *Emulation* – is the most demanding, and we should probably not expect it to be fully implemented anytime soon. It involves two steps. Firstly, we would have to investigate whether human–human therapeutic relationships can be analyzed as consisting of several simpler components or active ingredients, e.g., empathy, trust, positive regard, goal cohesion, understanding, etc. (95, 117–119). Secondly, we would have to try and reconstruct these components (or their counterparts) in the human-machine interaction. At this point, it is very difficult to assess to what extent this strategy is feasible. Let us take empathy as an example. While some authors claim that it is possible to build machines or chatbots capable of empathizing with people (120–123), others (124) point out that such optimism results from using an excessively restrictive characterization of empathy, e.g., identifying empathy with “empathic behavior.” But if, while developing artificial agents, we focus only on the behavioral aspect of empathy, neglecting all other aspects, such as emotional and phenomenal, maybe we are just developing more sophisticated forms of mimicry, and thus strategy three collapses to strategy two.

Finally, it might turn out to be the case that the most important active ingredient of the therapeutic relationship is something much easier to emulate or deliver by a CAI, e.g., autonomy support (125–127). According to Zuroff et al. (127), clients are “autonomously motivated when they experience themselves as having freely chosen their goals and the choice is felt to emanate from themselves” [p. 137]. Zuroff and colleagues suggest that such understood autonomy is the common factor of most efficient therapeutic interventions, and the main task of a therapist is to support rather than undermine it (128). If this is the case, CAIs may be well positioned to emulate it, simply by creating a context in which clients work through their problems without the impression of being dependent on, paternalized by, or forced into something they did not choose by an over-imposing therapist. We are yet to see whether, and in what ways, the technological solutions of the next decades will enable us to genuinely pursue the emulation strategy of solving The Problem of a Non-human Therapist.

5. The problem of a narrowly intelligent therapist

Based on the analysis of multiple available definitions, Shane Legg and Marcus Hutter (129) propose the following working definition of intelligence for AI research: “Intelligence measures an agent's ability to achieve goals in a wide range of environments” [p. 402]. The kind of artificial intelligence available now can achieve goals only in a relatively constrained or “narrow” range of environments (or solve only a constrained and “narrow” range of tasks), therefore it is typically called *artificial narrow intelligence* (ANI). In recent years, the use of ANI flooded different domains of our everyday lives. It is responsible for the accuracy of our Google searches, it crushes us at chess, it transforms speech to text, it drives autonomous vehicles, and so on.

Available ANIs are becoming able to deal with more and more complex environments. Some of the most important milestones in the recent ANI development were IBM's computer *Deep Blue* defeating the world champion in chess, and *AlphaGo* defeating (19 years later) the world champion in Go. Go is significantly more difficult for a computer to master than chess due to a gargantuan number of next possible move in any given position (130). *AlphaGo* achieved this goal, thanks to the use of the so-called *deep reinforcement learning* which combines reinforcement learning (a method of trial-and-error learning guided by reward maximization) with the use of deep neural networks (artificial neural networks using one or more hidden layers, and thus much more accurate than so-called *shallow* neural networks) (131). Even though narrow artificial intelligence is gradually becoming “less narrow,” we are still waiting for the so called artificial general intelligence (AGI), i.e., an artificial system capable of applying its intelligence to a virtually unrestricted range of tasks and environments, including ones that are new to it. This flexibility of intelligent thinking is the hallmark of human intelligence; therefore, AGI is also often referred to as *human-like AI*. While some argue that the path to AGI is relatively straightforward and we should expect to achieve this milestone within the next couple of years or decades (132), others are pessimistic about our prospects of ever building generally intelligent artificial systems (130) (for the results of an expert survey regarding

this issue see (133)).⁵ Be it as it may, at this point, we are stuck with ANI, and thus the last problem on the way towards AI-based psychotherapy is the following:

The Problem of a Narrowly Intelligent Therapist: Can a narrowly intelligent agent conduct psychotherapy?

Imagine a complicated social game in which you are supposed to coordinate your actions with your partner in such a way that you achieve a common goal that you have previously established. To win this game, you must excel in a number of supplementary tasks (or mini games). For example, you must accurately comprehend what your partner is saying; recognize cognitive errors they are making and suggest course correction; adequately read and react to their emotions, etc. The list of mini games contributing to the success in the big game is long but closed. There will be new scenarios but no new games along the way. Finally, each of the mini games has a well-defined set of rules, and at each point you will know whether you are doing well or failing.

Solving The Problem of a Narrowly Intelligent Therapist would require assessing whether psychotherapy can be construed as such a complicated social game. Is there a long but closed list of specific tasks, mastery of which enables one to conduct psychotherapy? Can we expect narrow AI to achieve mastery of each of these tasks? We are yet to see what the answers are to these questions. One benefit of posing this problem is that it sheds light on an issue which we typically neglect in the context of training human therapists, namely, that they possess many (if not most) skills necessary to conduct psychotherapy just by virtue of successfully participating in everyday social interactions.

The second aspect of the problem discussed in this section cuts even deeper into the roots of theoretical reflection about psychotherapy. One way of thinking about talk-therapy is to think about it as a series of conversations spanning across multiple meetings. Now, here is a question: “what are these conversations about?” An answer that comes to mind is that the conversations are about whatever the client and the therapist find relevant to the client’s suffering and whatever is worth touching upon to alleviate this suffering. This might remind us of the famous first sentence of Leo Tolstoy’s novel *Anna Karenina*: “Happy families are all alike; every unhappy family is unhappy in its own way.” Just as unhappiness of unhappy families, human mental suffering comes in a myriad of different forms. Does this mean that an agent capable of conducting fully fledged psychotherapy would have to be able to engage in a meaningful conversation about all of them? If yes, we probably should not expect an ANI to be able to do it.⁶

5 Notably, even the recent progress of AI technologies triggered by the development of LLMs does not guarantee a rapid arrival of AGI. As stated on the website of OpenAI — the developer of GPT-4 — “AGI could happen soon or far in the future; the takeoff speed from the initial AGI to more powerful successor systems could be slow or fast” (134).

6 A related question that we cannot discuss here in detail is whether interaction with an ANI can be properly characterized as a “conversation” (see (135)).

More specifically, from the technical point of view, the task of designing a dialogue system based on ANI forces one to confront the following dilemma:

The dialogue system dilemma: do we want our system to be general-purpose but its interaction with a user uncontrollable and unpredictable, or do we want the interaction to be predictable and controllable but restricted to fulfilling a specific, narrowly defined task?

Dialogue systems based on Large Language Models, such as ChatGPT, are general-purpose. We can equally well ask them to write a carrot cake recipe, describe the history of aviation, or enumerate the species of venomous snakes living in North America.⁷ At the same time, designers of such systems have no control over the structure of the interaction – the way it unfolds depends solely on users’ prompts. On the other hand, chatbots used in fields like commerce are designed for a specific purpose. A virtual assistant on an airline website will not be able to answer questions about baking recipes or history trivia, but it will guide its user through the process of achieving a specific goal, e.g., buying a plane ticket. Many such chatbots use a predefined structure or “frame” to guide the interaction (thus their name: “frame-based” (137)). A frame-based dialogue system used on an airline’s website may fulfill the objective of acquiring all information necessary to make a reservation by filling up a virtual form with slots for TIME OF FLIGHT, DESTINATION, NUMBER OF PASSENGERS, etc. Even though they allow some flexibility (e.g., if a user starts with “I’d like to fly on Wednesday,” the chatbot might reply: “Ok, where would you like to fly on Wednesday” and if a user starts with “I’d like to fly to Toronto” the chatbot may reply with: “Ok, when would you like to fly to Toronto?”), the goal of the interaction is fixed, predefined, and narrow.

Contemporary mental health chatbots are precisely such simple frame-based systems guiding users through predesigned steps of built-in therapeutic interventions or exercises (27, 34, 138). Whenever there is a place for unscripted, natural language input from the user, a chatbot has to classify it as falling into one of the coarse-grained, pre-defined categories.⁸ It severely limits the chatbot’s ability to appropriately react to the user’s input, as witnessed, for example, by the following exchange:

- [1] Woebot: Okay, want to tell me more about what’s going on?
- [2] User: I’m feeling weird. It’s like I’m empty inside. When I look at the faces of my friends and family, I feel like they are empty as well. That’s why we cannot connect with each other. There’s nothing to hold on to.
- [3] Woebot: Sounds like you are dealing with feeling low...

7 Nevertheless, chatbots based on LLMs are currently still helpless when confronted with sophisticated metaphors and absurdities (136) or even certain trivial logical puzzles (35).

8 This process is called Named Entity Recognition. The “entities” it recognizes are “specific information that is extracted from the user’s input that maps the natural language phrases with the canonical phrases to understand the intent.” (27) [p. 3759].

In the future, we should expect chatbots to be able to recognize a much broader range of topics and intents brought up by users. Most likely, they will also be equipped with functionalities allowing them to learn how to respond more appropriately by learning from a user's earlier inputs (27) and maybe even adapt to the user's personality and thus increase their engagement (114). However, as long as we would like our ANIs to primarily implement specific therapeutic techniques, we would also have to keep the range of the topics and tasks they can engage in strictly restricted. Therefore, The Problem of the Narrowly Intelligent Therapist remains open. In light of it, we may be forced to admit either that artificial therapists are impossible to construe until we reach the level of technological engagement equal to AGI or that "psychotherapy" in the future will mean something different than it means today.

6. Conclusion

The use of AI is becoming increasingly widespread in the field of mental health care. In particular, the first promising attempts are being made to design AI-based technologies capable of providing psychotherapeutic help. While the available research demonstrates that mental health chatbots can, undoubtedly, be very helpful to at least some of their users, the goal of this paper was to outline the scope of the challenge of developing fully fledged AI-based psychotherapy. We offered this outline in the form of three major problems which have to be faced before we will be able to schedule our first appointments with artificial therapists. We find it very likely that in the future, each of these challenges will be overcome in one way or another. Until then, however, it is crucial to be honest and explicit about the limited role an AI can play in psychotherapeutic processes.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

References

- Rehm J, Shield KD. Global burden of disease and the impact of mental and addictive disorders. *Curr Psychiatry Rep.* (2019) 21:10. doi: 10.1007/s11920-019-0997-0
- Steel Z, Marnane C, Iranpour C, Chey T, Jackson JW, Patel V, et al. The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013. *Int J Epidemiol.* (2014) 43:476–93. doi: 10.1093/ije/dyu038
- Xiong J, Lipsitz O, Nasri F, Lui LMW, Gill H, Phan L, et al. Impact of COVID-19 pandemic on mental health in the general population: a systematic review. *J Affect Disord.* (2020) 277:55–64. doi: 10.1016/j.jad.2020.08.001
- Patel V, Saxena S, Lund C, Thornicroft G, Baingana F, Bolton P, et al. The lancet commission on global mental health and sustainable development. *Lancet.* (2018) 392:1553–98. doi: 10.1016/S0140-6736(18)31612-X
- D'Alfonso S. AI in mental health. *Curr Opin Psychol.* (2020) 36:112–7. doi: 10.1016/j.copsyc.2020.04.005
- Fiske A, Henningsen P, Buyx A. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *J Med Internet Res.* (2019) 21:e13216. doi: 10.2196/13216
- Graham S, Depp C, Lee EE, Nebeker C, Tu X, Kim H-C, et al. Artificial intelligence for mental health and mental illnesses: an overview. *Curr Psychiatr Rep.* (2019) 21:116. doi: 10.1007/s11920-019-1094-0
- Haque A, Guo M, Miner AS, Fei-Fei L. Measuring depression symptom severity from spoken language and 3D facial expressions. (2018) Available at: <http://arxiv.org/abs/1811.08592> (Accessed July 20, 2022).
- Mastoras R-E, Iakovakis D, Hadjidimitriou S, Charisis V, Kassie S, Alsaadi T, et al. Touchscreen typing pattern analysis for remote detection of the depressive tendency. *Sci Rep.* (2019) 9:13414. doi: 10.1038/s41598-019-50002-9
- Ware S, Yue C, Morillo R, Lu J, Shang C, Bi J, et al. Predicting depressive symptoms using smartphone data. *Smart Health.* (2020) 15:100093. doi: 10.1016/j.smhl.2019.100093
- Corcoran CM, Carrillo F, Fernández-Slezak D, Bedi G, Klim C, Javitt DC, et al. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatr.* (2018) 17:67–75. doi: 10.1002/wps.20491
- Dwyer DB, Cabral C, Kambeitz-Ilankovic L, Sanfelici R, Kambeitz J, Calhoun V, et al. Brain subtyping enhances the neuroanatomical discrimination of schizophrenia. *Schizophr Bull.* (2018) 44:1060–9. doi: 10.1093/schbul/sby008
- Iter D, Yoon J, Jurafsky D. Automatic detection of incoherent speech for diagnosing schizophrenia. *Proceedings of the 5th workshop on computational linguistics and clinical psychology: From keyboard to clinic.* New Orleans, LA: Association for Computational Linguistics (2018). p. 136–146
- Chekroud AM, Bondar J, Delgado J, Doherty G, Wasil A, Fokkema M, et al. The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatr.* (2021) 20:154–70. doi: 10.1002/wps.20882
- Thieme A, Belgrave D, Doherty G. Machine learning in mental health: a systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM Trans Comput Hum Interact.* (2020) 27:1–53. doi: 10.1145/3398069

Author contributions

JPG and MH reviewed the literature, critically discussed the theoretical stance, commented on the manuscript, and approved the submitted version of the manuscript. JPG wrote the draft of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

JPG research and open access publication were supported by a grant from the Priority Research Area 'Society of the Future' under the Strategic Programme 'Excellence Initiative' at Jagiellonian University. MH research was supported by the National Science Centre, Poland (grant number: 2021/43/B/HS1/02868).

Acknowledgments

We would like to thank Grzegorz Gaszczyk, Kinga Wołoszyn, and two reviewers for their thought-provoking comments and suggestions.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

16. van Breda W. Predictive modeling in E-mental health: Exploring applicability in personalised depression treatment. (PhD-Thesis-Research and graduation internal). Amsterdam: Vrije Universiteit Amsterdam (2020). Available at: <https://research.vu.nl/ws/portalfiles/portal/101278982/205414.pdf> (Accessed March 1, 2023)
17. van Breda W, Bremer V, Becker D, Hoogendoorn M, Funk B, Ruwaard J, et al. Predicting therapy success for treatment as usual and blended treatment in the domain of depression. *Internet Inter.* (2018) 12:100–4. doi: 10.1016/j.invent.2017.08.003
18. Huijnen CAGJ, Lexis MAS, Jansens R, de Witte LP. How to implement robots in interventions for children with autism? A co-creation study involving people with autism, parents, and professionals. *J Autism Dev Disord.* (2017) 47:3079–96. doi: 10.1007/s10803-017-3235-9
19. Góngora Alonso S, Hamrioui S, de la Torre DI, Motta Cruz E, López-Coronado M, Franco M. Social robots for people with aging and dementia: a systematic review of literature. *Telemed E-Health.* (2019) 25:533–40. doi: 10.1089/tmj.2018.0051
20. Craig TK, Rus-Calafell M, Ward T, Leff JB, Huckvale M, Howarth E, et al. AVATAR therapy for auditory verbal hallucinations in people with psychosis: a single-blind, randomised controlled trial. *Lancet Psychiatr.* (2018) 5:31–40. doi: 10.1016/S2215-0366(17)30427-3
21. Dellazizzo L, Potvin S, Phraxayavong K, Lalonde P, Dumais A. Avatar therapy for persistent auditory verbal hallucinations in an ultra-resistant schizophrenia patient: a case report. *Front Psych.* (2018) 9:131. doi: 10.3389/fpsy.2018.00131
22. Boucher EM, Harake NR, Ward HE, Stoeckl SE, Vargas J, Minkel J, et al. Artificially intelligent chatbots in digital mental health interventions: a review. *Expert Rev Med Devices.* (2021) 18:37–49. doi: 10.1080/17434440.2021.2013200
23. Brown JEH, Halpern J. AI chatbots cannot replace human interactions in the pursuit of more inclusive mental healthcare. *SSM-Ment Health.* (2021) 1:100017. doi: 10.1016/j.ssmmh.2021.100017
24. Gaffney H, Mansell W, Tai S. Conversational agents in the treatment of mental health problems: mixed-method systematic review. *JMIR Ment Health.* (2019) 6:e14166. doi: 10.2196/14166
25. Huston B. Could a robot be your psychotherapist? (2020). Available at: https://digitalcommons.du.edu/capstone_masters/378 (Accessed April 23, 2023).
26. Miner AS, Shah N, Bullock KD, Arnow BA, Bailenson J, Hancock J. Key considerations for incorporating conversational AI in psychotherapy. *Front Psych.* (2019) 10:746. doi: 10.3389/fpsy.2019.00746
27. Pandey S, Sharma S, Wazir S. Mental healthcare chatbot based on natural language processing and deep learning approaches: ted the therapist. *Int J Inf Technol.* (2022) 14:3757–66. doi: 10.1007/s41870-022-00999-6
28. Sedlakova J, Trachsel M. Conversational artificial intelligence in psychotherapy: a new therapeutic tool or agent? *Am J Bioeth.* (2022) 23:4–13. doi: 10.1080/15265161.2022.2048739
29. Tekin Ş. Is big data the new stethoscope? Perils of digital phenotyping to address mental illness. *Philos Technol.* (2021) 34:447–61. doi: 10.1007/s13347-020-00395-7
30. Tekin Ş. Ethical issues surrounding artificial intelligence technologies in mental health: psychotherapy chatbots. In: G Robson and J Tsou, Eds. *Technology ethics: A philosophical introduction and readings.* Routledge (2023).
31. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *Can J Psychiatr.* (2019) 64:456–64. doi: 10.1177/0706743719828977
32. American Psychological Association. What is psychotherapy? (2017) Available at: <https://www.apa.org/ptsd-guideline/patients-and-families/psychotherapy> (Accessed July 20, 2022)
33. Norcross JC. An eclectic definition of psychotherapy. In: JK Zeig and WM Munion, Eds. *What is psychotherapy?*. San Francisco: Jossey-Bass (1990).
34. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. *J Am Med Inform Assoc.* (2018) 25:1248–58. doi: 10.1093/jamia/ocy072
35. Floridi L. AI as agency without intelligence: on ChatGPT, large language models, and other generative models. *Philos Technol.* (2023) 36. doi: 10.1007/s13347-023-00621-y
36. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. *Chain-of-thought prompting elicits reasoning in large language models.* (2023).
37. Kosinski M. *Theory of mind may have spontaneously emerged in large language models.* (2023).
38. Ullman T. Large language models fail on trivial alterations to theory-of-mind tasks. (2023) Available at: <http://arxiv.org/abs/2302.08399> (Accessed April 23, 2023).
39. Emerson M, Torous J. How to choose a mental health app | psyche guides. Psyche (Stuttgart) (2022) Available at: <https://psyche.co/guides/how-to-choose-a-mental-health-app-that-can-actually-help> (Accessed July 20, 2022).
40. Abd-alrazaq AA, Alajlani M, Alalwan AA, Bewick BM, Gardner P, Househ M. An overview of the features of chatbots in mental health: a scoping review. *Int J Med Inf.* (2019) 132:103978. doi: 10.1016/j.ijmedinf.2019.103978
41. Woebot Health. Woebot health (2022) Available at: <https://woebothealth.com/> (Accessed July 7, 2022).
42. Clark DA. “Cognitive restructuring,” in *The Wiley handbook of cognitive behavioral therapy.* eds. S. G. Hofmann and D. Dozois. John Wiley & Sons, Ltd. (2014). Vol 1, 23–44.
43. Blandford A. HCI for health and wellbeing: challenges and opportunities. *Int J Hum-Comput Stud.* (2019) 131:41–51. doi: 10.1016/j.ijhcs.2019.06.007
44. Luxton DD. Ethical implications of conversational agents in global public health. *Bull World Health Organ.* (2020) 98:285–7. doi: 10.2471/BLT.19.237636
45. Manriquez Roa T, Biller-Andorno N, Trachsel M. The ethics of artificial intelligence in psychotherapy In: M Trachsel, Ş Tekin, N Biller-Andorno, J Gaab and JZ Sadler, editors. *The Oxford handbook of psychotherapy ethics.* Oxford: Oxford University Press (2021). 744–58.
46. Nurgalieva L, Doherty G. Privacy and security in digital therapeutics In: N Jacobson, T Kowatsch and L Marsch, editors. *Digital therapeutics for mental health and addiction:* Academic Press (2023). 189–204.
47. Wysa—Everyday Mental Health. Wysa—everyday Ment health (2022) Available at: <https://www.wysa.io/> (Accessed July 7, 2022).
48. Mental Health Chatbot | X2. (2022) Available at: <https://www.x2ai.com/individuals> (Accessed July 20, 2022).
49. Abd-Alrazaq AA, Rababeh A, Alajlani M, Bewick BM, Househ M. Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. *J Med Internet Res.* (2020) 22:e16021. doi: 10.2196/16021
50. Goldberg SB, Lam SU, Simonsson O, Torous J, Sun S. Mobile phone-based interventions for mental health: a systematic meta-review of 14 meta-analyses of randomized controlled trials. *PLOS Digit Health.* (2022) 1:e0000002. doi: 10.1371/journal.pdig.0000002
51. He Y, Yang L, Zhu X, Wu B, Zhang S, Qian C, et al. Mental health Chatbot for young adults with depressive symptoms during the COVID-19 pandemic: single-blind, three-arm randomized controlled trial. *J Med Internet Res.* (2022) 24:e40719. doi: 10.2196/40719
52. Liu H, Peng H, Song X, Xu C, Zhang M. Using AI chatbots to provide self-help depression interventions for university students: a randomized trial of effectiveness. *Internet Interv.* (2022) 27:100495. doi: 10.1016/j.invent.2022.100495
53. Lim SM, Shiau CWC, Cheng LJ, Lau Y. Chatbot-delivered psychotherapy for adults with depressive and anxiety symptoms: a systematic review and meta-regression. *Behav Ther.* (2022) 53:334–47. doi: 10.1016/j.beth.2021.09.007
54. Prochaska JJ, Vogel EA, Chieng A, Baiocchi M, Maglalang DD, Pajarito S, et al. A randomized controlled trial of a therapeutic relational agent for reducing substance misuse during the COVID-19 pandemic. *Drug Alcohol Depend.* (2021) 227:108986. doi: 10.1016/j.drugalcdep.2021.108986
55. Prochaska JJ, Vogel EA, Chieng A, Kendra M, Baiocchi M, Pajarito S, et al. A therapeutic relational agent for reducing problematic substance use (Woebot): development and usability study. *J Med Internet Res.* (2021) 23:e24850. doi: 10.2196/24850
56. Suharwardy S, Ramachandran M, Leonard SA, Gunaseelan A, Lyell DJ, Darcy A, et al. Feasibility and impact of a mental health chatbot on postpartum mental health: a randomized controlled trial. *AJOG Glob Rep.* (2023) 2023:100165. doi: 10.1016/j.xagr.2023.100165
57. He Y, Yang L, Li T, Qian C, Su Z, Zhang Q. Conversational agent interventions for mental health problems: systematic review and meta-analysis of randomized controlled trials. *J Med Internet Res.* (2023) 25:e43862. doi: 10.2196/43862
58. Grodniewicz JB, Hohol M. Therapeutic conversational artificial intelligence and the Acquisition of Self-understanding. *Am J Bioeth.* (2023) 23:59–61. doi: 10.1080/15265161.2023.2191021
59. Lacey M. Psychodynamic psychotherapy, insight, and therapeutic action. *Clin Psychol Sci Pract.* (2014) 21:154–71. doi: 10.1111/cpsp.12065
60. Prochaska JO, Norcross JC. *Systems of psychotherapy: a transtheoretical analysis.* 9th ed. New York, NY: Oxford University Press (2018).
61. Cook JM, Biyanova T, Elhai J, Schnurr PP, Coyne JC. What do psychotherapists really do in practice? An internet study of over 2,000 practitioners. *Psychother Theory Res Pract Train.* (2010) 47:260–7. doi: 10.1037/a0019788
62. Arkowitz H, Messer S Eds. *Psychoanalytic therapy and behavior therapy: Is integration possible?* New York, NY: Springer (1984).
63. Poznanski JJ, McLennan J. Conceptualizing and measuring counselors’ theoretical orientation. *J Couns Psychol.* (1995) 42:411–22. doi: 10.1037/0022-0167.42.4.411
64. Rosenzweig S. Some implicit common factors in diverse methods of psychotherapy. *Am J Orthopsychiatry.* (1936) 6:412–5. doi: 10.1111/j.1939-0025.1936.tb05248.x
65. Beutler LE. “Eclectic psychotherapy,” in *Encyclopedia of psychology.* eds. A. E. Kazdin. Oxford University Press (2000). Vol 3, 128–129.
66. Frank JD, Frank JB. *Persuasion and healing: A comparative study of psychotherapy,* vol. 18. 3rd ed. Baltimore, MD, US: Johns Hopkins University Press (1991).
67. Goldfried MR. Toward the delineation of therapeutic change principles. *Am Psychol.* (1980) 35:991–9. doi: 10.1037//0003-066x.35.11.991

68. Karasu TB. The specificity versus nonspecificity dilemma: toward identifying therapeutic change agents. *Am J Psychiatry*. (1986) 143:687–95. doi: 10.1176/ajp.143.6.687
69. Kleinke CL. *Common principles of psychotherapy*, vol. 20. Belmont, CA, US: Thomson Brooks/Cole Publishing Co (1994).
70. Wampold BE, Imel ZE. *The great psychotherapy debate: the evidence for what makes psychotherapy work*. 2nd ed. New York: Routledge (2015).
71. Weinberger J. Common factors aren't so common: the common factors dilemma. *Clin Psychol Sci Pract*. (1995) 2:45–69. doi: 10.1111/j.1468-2850.1995.tb00024.x
72. Buckman JR, Barker C. Therapeutic orientation preferences in trainee clinical psychologists: personality or training? *Psychother Res*. (2010) 20:247–58. doi: 10.1080/10503300903352693
73. Heffler B, Sandell R. The role of learning style in choosing one's therapeutic orientation. *Psychother Res J Soc Psychother Res*. (2009) 19:283–92. doi: 10.1080/10503300902806673
74. Tartakovsky E. The motivational foundations of different therapeutic orientations as indicated by therapists' value preferences. *Psychother Res*. (2016) 26:352–64. doi: 10.1080/10503307.2014.989289
75. Blackwell A. Artificial intelligence meets mental health therapy. TED Talk TEDxNatick (2020) Available at: https://www.ted.com/talks/andy_blackwell_artificial_intelligence_meets_mental_health_therapy (Accessed February 10, 2023)
76. Becker KD, Chorpita BF, Daleiden EL. Improvement in symptoms versus functioning: how do our best treatments measure up? *Adm Policy Ment Health Ment Health Serv Res*. (2011) 38:440–58. doi: 10.1007/s10488-010-0332-x
77. Norcross JC. A primer on psychotherapy integration In: . *Handbook of psychotherapy integration*, Oxford series in clinical psychology. 2nd ed. New York, NY, US: Oxford University Press (2005). 3–23.
78. Gaffney H, Mansell W, Tai S. Agents of change: understanding the therapeutic processes associated with the helpfulness of therapy for mental health problems with relational agent MYLO. *Digit Health*. (2020) 6:20520762091158. doi: 10.1177/2055207620911580
79. Carey TA. *The method of levels: how to do psychotherapy without getting in the way* Living Control Systems Publishing (2006).
80. Carey TA, Mansell W, Tai S. A biopsychosocial model based on negative feedback and control. *Front Hum Neurosci*. (2014) 8:94. doi: 10.3389/fnhum.2014.00094
81. Powers WT. *Behavior: The control of perception*. New York: Hawthorne (1973). 296 p.
82. Mansell W. Method of levels: is it the most parsimonious psychological therapy available? *Rev Psicoter*. (2018) 29:135–43.
83. Bennion MR, Hardy GE, Moore RK, Kellett S, Millings A. Usability, acceptability, and effectiveness of web-based conversational agents to facilitate problem solving in older adults: controlled study. *J Med Internet Res*. (2020) 22:e16794. doi: 10.2196/16794
84. Wrightson-Hester A-R, Anderson G, Dunstan J, McEvoy P, Sutton C, Myers B, et al. *Manage your life online (MYLO): Co-design and case-series of an artificial therapist to support youth mental health* (2023) doi: 10.31234/osf.io/zjw8p
85. Society of Clinical Psychology. *Treatments* (2023) Available at: <https://div12.org/treatments/> (Accessed May 7, 2022)
86. Ewbank MP, Cummins R, Tablan V, Bateup S, Catarino A, et al. Quantifying the association between psychotherapy content and clinical outcomes using deep learning. *JAMA Psychiat*. (2020) 77:35–43. doi: 10.1001/jamapsychiatry.2019.2664
87. Fitzpatrick KK, Darcy A, Vierhile M. 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 Ment Health*. (2017) 4:e7785:e19. doi: 10.2196/mental.7785
88. Murphy ST, Garcia RA, Cheavens JS, Strunk DR. The therapeutic alliance and dropout in cognitive behavioral therapy of depression. *Psychother Res*. (2022) 32:995–1002. doi: 10.1080/10503307.2021.2025277
89. Wilmots E, Midgley N, Thackeray L, Reynolds S, Loades M. The therapeutic relationship in cognitive behaviour therapy with depressed adolescents: a qualitative study of good-outcome cases. *Psychol Psychother*. (2020) 93:276–91. doi: 10.1111/papt.12232
90. Beck AT, Rush AJ, Shaw BF, Emery G eds. *Cognitive therapy of depression*. New York: Guilford Press (1979).
91. Baker LR. *Persons and bodies: a constitution view*. Cambridge, U.K.; New York: Cambridge University Press (2000).
92. Schlosser M. Agency, In: EN Zalta, Ed. *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University (2019) Available at: <https://plato.stanford.edu/archives/win2019/entries/agency/>
93. Floridi L, Sanders JW. On the morality of artificial agents. *Minds Mach*. (2004) 14:349–79. doi: 10.1023/B:MIND.0000035461.63578.9d
94. Askjer S, Mathiasen K. The working alliance in blended versus face-to-face cognitive therapy for depression: a secondary analysis of a randomized controlled trial. *Internet Interv*. (2021) 25:100404. doi: 10.1016/j.invent.2021.100404
95. Norcross JC, Lambert MJ. (Eds.) "Evidence-based psychotherapy relationship: the third task force," in *Psychotherapy relationships that work*: Oxford University Press (2019). Vol. 1, 1–23.
96. Holohan M, Fiske A. "Like I'm talking to a real person": exploring the meaning of transference for the use and design of AI-based applications in psychotherapy. *Front Psychol*. (2021) 12:720476. doi: 10.3389/fpsyg.2021.720476
97. World Health Organization. *Mental health atlas 2020*. Geneva: World Health Organization (2021). Available at: <https://apps.who.int/iris/handle/10665/345946> (Accessed July 16, 2022)
98. Henderson C, Evans-Lacko S, Thornicroft G. Mental illness stigma, help seeking, and public health programs. *Am J Public Health*. (2013) 103:777–80. doi: 10.2105/AJPH.2012.301056
99. Allen SF, Wetherell MA, Smith MA. Online writing about positive life experiences reduces depression and perceived stress reactivity in socially inhibited individuals. *Psychiatry Res*. (2020) 284:112697. doi: 10.1016/j.psychres.2019.112697
100. Carey TA, Kelly RE, Mansell W, Tai SJ. What's therapeutic about the therapeutic relationship? A hypothesis for practice informed by perceptual control theory. *Cogn Behav Ther*. (2012) 5:47–59. doi: 10.1017/S1754470X12000037
101. Bordin ES. The generalizability of the psychoanalytic concept of the working alliance. *Psychother Theory Res Pract*. (1979) 16:252–60. doi: 10.1037/h0085885
102. Kaveladze B, Schueller SM. A digital therapeutic alliance in digital mental health In: N Jacobson, T Kowatsch and L Marsch, editors. *Digital therapeutics for mental health and addiction*: Academic Press (2023). 87–98.
103. Lederman R, D'Alfonso S. The digital therapeutic Alliance: prospects and considerations. *JMIR Ment Health*. (2021) 8:e31385. doi: 10.2196/31385
104. Berry K, Salter A, Morris R, James S, Bucci S. Assessing therapeutic alliance in the context of mHealth interventions for mental health problems: development of the Mobile Agnew relationship measure (mARM) questionnaire. *J Med Internet Res*. (2018) 20:e90. doi: 10.2196/jmir.8252
105. Henson P, Peck P, Torous J. Considering the therapeutic alliance in digital mental health interventions. *Harv Rev Psychiatry*. (2019) 27:268–73. doi: 10.1097/HRP.0000000000000224
106. Henson P, Wisniewski H, Hollis C, Keshavan M, Torous J. Digital mental health apps and the therapeutic alliance: initial review. *BJPsych Open*. (2019) 5:e15. doi: 10.1192/bjo.2018.86
107. Beatty C, Malik T, Meheli S, Sinha C. Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): a mixed-methods study. *Front Digit Health*. (2022) 4:847991. doi: 10.3389/fdgth.2022.847991
108. D'Alfonso S, Lederman R, Bucci S, Berry K. The digital therapeutic alliance and human-computer interaction. *JMIR Ment Health*. (2020) 7:e21895. doi: 10.2196/21895
109. Darcy A, Daniels J, Salinger D, Wicks P, Robinson A. Evidence of human-level bonds established with a digital conversational agent: cross-sectional, retrospective observational study. *J Med Internet Res*. (2021) 5:e27868. doi: 10.2196/27868
110. Dosovitsky G, Bunge EL. Bonding with bot: user feedback on a Chatbot for social isolation. *Front Digit Health*. (2021) 3:735053. doi: 10.3389/fdgth.2021.735053
111. Hauser-Ulrich S, Künzli H, Meier-Peterhans D, Kowatsch T. A smartphone-based health care Chatbot to promote self-management of Chronic Pain (SELMA): pilot randomized controlled trial. *JMIR Mhealth Uhealth*. (2020) 8:e15806. doi: 10.2196/15806
112. Tremain H, McEnery C, Fletcher K, Murray G. The therapeutic alliance in digital mental health interventions for serious mental illnesses: narrative review. *JMIR Ment Health*. (2020) 7:e17204. doi: 10.2196/17204
113. Abd-Alrazaq AA, Alajlani M, Ali N, Denecke K, Bewick BM, Househ M. Perceptions and opinions of patients about mental health Chatbots: scoping review. *J Med Internet Res*. (2021) 23:e17828. doi: 10.2196/17828
114. Ahmad R, Siemon D, Gnewuch U, Robra-Bissantz S. Designing personality-adaptive conversational agents for mental health care. *Inf Syst Front*. (2022) 24:923–43. doi: 10.1007/s10796-022-10254-9
115. Nißen M, Rügger D, Stieger M, Flückiger C, Allemann M, van Wangenheim F, et al. The effects of health care Chatbot personas with different social roles on the client-Chatbot bond and usage intentions: development of a design codebook and web-based study. *J Med Internet Res*. (2022) 24:e32630. doi: 10.2196/32630
116. Weizenbaum J. *Computer power and human reason: from judgment to calculation*. San Francisco: W. H. Freeman (1976).
117. Elliott R, Bohart AC, Watson JC, Murphy D. "Empathy," in *Psychotherapy relationships that work*. eds. J. C. Norcross and M. J. Lambert. Oxford University Press (2019). 245–87.
118. Farber BA, Suzuki JY, Lynch DA. "Positive regard and affirmation," in *Psychotherapy relationships that work*. eds. J. C. Norcross and M. J. Lambert. Oxford University Press (2019). 288–322.
119. Flückiger JC, Del Re AC, Wampold BE, Horvath AO. "Alliance in adult psychotherapy," in *Psychotherapy relationships that work*. eds. J. C. Norcross and M. J. Lambert. Oxford University Press (2019). 24–78.
120. Kozima H, Nakagawa C, Yano H. Can a robot empathize with people? *Artif Life Robot*. (2004) 8:83–8. doi: 10.1007/s10015-004-0293-9

121. Leite I, Castellano G, Pereira A, Martinho C, Paiva A. Empathic robots for long-term interaction. *Int J Soc Robot.* (2014) 6:329–41. doi: 10.1007/s12369-014-0227-1
122. Skjuve M, Følstad A, Fostervold KI, Brandtzaeg PB. My Chatbot companion—a study of human-Chatbot relationships. *Int J Hum-Comput Stud.* (2021) 149:102601. doi: 10.1016/j.ijhcs.2021.102601
123. Skjuve M, Følstad A, Fostervold KI, Brandtzaeg PB. A longitudinal study of human–chatbot relationships. *Int J Hum Comput Stud.* (2022) 168:102903. doi: 10.1016/j.ijhcs.2022.102903
124. Malinowska JK. What does it mean to empathise with a robot? *Minds Mach.* (2021) 31:361–76. doi: 10.1007/s11023-021-09558-7
125. Buchholz JL, Abramowitz JS. The therapeutic alliance in exposure therapy for anxiety-related disorders: a critical review. *J Anxiety Disord.* (2020) 70:102194. doi: 10.1016/j.janxdis.2020.102194
126. Deci EL, Ryan RM. The “what” and “why” of goal pursuits: human needs and the self-determination of behavior. *Psychol Inq.* (2000) 11:227–68. doi: 10.1207/S15327965PLI1104_01
127. Zuroff DC, Koestner R, Moskowitz DS, McBride C, Marshall M, Bagby MR. Autonomous motivation for therapy: a new common factor in brief treatments for depression. *Psychother Res.* (2007) 17:137–47. doi: 10.1080/10503300600919380
128. Markland D, Ryan RM, Tobin VJ, Rollnick S. Motivational interviewing and self-determination theory. *J Soc Clin Psychol.* (2005) 24:811–31. doi: 10.1521/jscp.2005.24.6.811
129. Legg S, Hutter M. A collection of definitions of intelligence. (2007) Available at: <http://arxiv.org/abs/0706.3639> (Accessed July 19, 2022).
130. Fjelland R. Why general artificial intelligence will not be realized. *Humanit Soc Sci Commun.* (2020) 7:10. doi: 10.1057/s41599-020-0494-4
131. Li Y. Deep reinforcement learning. (2018) Available at: <http://arxiv.org/abs/1810.06339> (Accessed July 19, 2022).
132. Silver D, Singh S, Precup D, Sutton RS. Reward is enough. *Artif Intell.* (2021) 299:103535. doi: 10.1016/j.artint.2021.103535
133. 2022 expert survey on Progress in AI. AI Impacts (2022) Available at: <https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/> (Accessed April 24, 2023).
134. Altman S. Planning for AGI and beyond. (2023) Available at: <https://openai.com/blog/planning-for-agi-and-beyond> (Accessed April 24, 2023).
135. Mallory F. Fictionalism about chatbots. *Ergo.* (2023)
136. Hofstadter D. Artificial neural networks today are not conscious, according to Douglas Hofstadter. *De Economist* (2022) Available at: <https://www.economist.com/by-invitation/2022/06/09/artificial-neural-networks-today-are-not-conscious-according-to-douglas-hofstadter> (Accessed July 20, 2022).
137. Harms J-G, Kucherbaev P, Bozzon A, Houben G-J. Approaches for dialog management in conversational agents. *IEEE Internet Comput.* (2019) 23:13–22. doi: 10.1109/MIC.2018.2881519
138. Darcy A, Beaudette A, Chiauzzi E, Daniels J, Goodwin K, Mariano TY, et al. Anatomy of a Woebot[®] (WB001): agent guided CBT for women with postpartum depression. *Expert Rev Med Devices.* (2022) 19:287–301. doi: 10.1080/17434440.2022.2075726