A narrative review of the active ingredients in psychotherapy delivered by conversational agents

Arthur Bran Herbener, Michał Klincewicz, Malene Flensborg Damholdt

PII: S2451-9588(24)00034-4

DOI: https://doi.org/10.1016/j.chbr.2024.100401

Reference: CHBR 100401

To appear in: *Computers in Human Behavior Reports*

Received Date: 22 November 2023

Revised Date: 19 February 2024

Accepted Date: 1 March 2024

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

A Narrative Review of the Active Ingredients in Psychotherapy Delivered by Conversational Agents

Arthur Bran Herbener[1], Michał Klincewicz[2, 4], and Malene Flensborg Damholdt[3]

[1,3]Department of Psychology and Behavioral Sciences, Aarhus University
[2]Department of Cognitive Science and Artificial Intelligence, Tilburg University
[4] Department of Cognitive Science, Institute of Philosophy, Jagiellonian University

Authors note

We have no known conflict of interest to disclose. Correspondence concerning this article should be addressed to Arthur Bran Herbener, Department of Psychology and Behavioral Sciences, Aarhus University, Bartholins Allé 11, 8000, Aarhus C, Denmark. Email: abh@psy.au.dk

Abstract

The present narrative review seeks to unravel where we are now, and where we need to go to delineate the active ingredients in psychotherapy delivered by conversational agents (e.g., chatbots). While psychotherapy delivered by conversational agents has shown promising effectiveness for depression, anxiety, and psychological distress across several randomized controlled trials, little emphasis has been placed on the therapeutic processes in these interventions. The theoretical framework of this narrative review is grounded in prominent perspectives on the active ingredients in psychotherapy. Based on search terms derived from various theoretical perspectives, we conducted a systematic literature search of four scientific databases and identified 17 studies. Across the selected studies, three themes emerged: relationship variables, emotional venting, and cognitive factors. While methodological issues compromise the epistemic value of this evidence base, core questions also remain to be answered. Such questions include, but are not limited to, whether humans can form affectionate bonds to inanimate beings and whether these kind of mental health treatments should be understood as psychotherapy or something else. Researchers should therefore be cautious when applying theories of psychotherapy in the realm of conversational agents. We conclude the paper by introducing recommendations for future research, which we hope will help instigate methodologically sound studies in this field.

Keywords

Conversational agent, psychotherapy, active ingredient, artificial intelligence, chatbot, mental health.

1. Introduction

"It's the relationship that heals" stated American psychiatrist Irvin Yalom (1989, p. 122) regarding how psychotherapy works. New findings put this notion to the test – or at least question whether the relationship must be between humans. Recent years have featured an explosive upsurge in research on psychotherapy delivered by conversational agents, such as chatbots and social robots. A meta-analysis identified 32 randomized controlled trials of digital conversational agents delivering mental health interventions, demonstrating promising effectiveness in decreasing symptoms of depression, anxiety, and psychological distress in adults (He et al., 2023). With the reservation that this research field remains in its early stages, psychotherapy delivered by conversational agents seems to be a promising alternative to established mental health interventions, overcoming core treatment barriers such as economic costs, waiting lists, and geographical distances (Kazdin, 2018).

While there is a long tradition for studying how psychotherapy works (Cuijpers et al., 2019; Wampold & Imel, 2015), such a research direction remains to be established in the realm of conversational agents. In psychotherapy research, so-called *process research* partly concentrates on what aspects of therapy account for therapeutic effectiveness, also called the *active ingredients* (McAleavey & Castonguay, 2015). Process-research is important, as it helps identify effective and redundant treatment components – answering the question: what works for whom, and under which circumstances (Kazdin, 2007). However, theory and evidence from process-oriented psychotherapy research are not necessarily compatible with conversational agents. While conversational agents, to varying degrees, can engage in human-sounding conversations, the issue here is that there may be more than the literal meanings conveyed through spoken words that account for the effectiveness of therapy. This includes, but is not limited to, the interpersonal relationship and expectations of improvement arising from the sociocultural context surrounding therapy (Frank & Frank, 1991; Wampold & Imel, 2015). Therefore, it cannot automatically be assumed that CA-psychotherapy works through similar causal pathways as traditional psychotherapy.

Considering the promising meta-analytical evidence on CA-psychotherapy for emotional problems coupled with the limited knowledge of the active ingredients, the time has come to ask *how* CA-psychotherapy brings about emotional improvements in adults. To instigate a process-oriented research direction in the realm of conversational agents, we conducted a narrative review examining *where we are now* and *where we need to go* to expand our knowledge of the active ingredients in CA-psychotherapy. The analytical framework for the narrative review is grounded in different theoretical perspectives on active ingredients in traditional, human-delivered psychotherapy. Specifically, we systematically searched for research on this topic and delineated common themes that suggest candidates for the active ingredients in CA-psychotherapy. Based on our findings, we outline

recommendations for future research to instigate scientific discussions and endeavors into exploring the active ingredients in CA-psychotherapy.

2. Background

2.1. Psychotherapy delivered by conversational agents

Even though research on conversational agents has experienced notable growth in recent years, the idea itself is not a recent development. In the 1960s, software engineer Joseph Weizenbaum created the pioneering chatbot *Eliza*, which employed simple pattern-matching and rephrasing techniques to simulate a Rogerian therapist (Weizenbaum, 1966). Thenceforth, a variety of conversational agents has been developed. Before we turn to recent scientific endeavors into CA-psychotherapy, we briefly outline the conceptual features of conversational agents.

The term conversational agent is best understood as an umbrella construct that encompass a range of software systems developed to engage in verbal conversations with humans. Specifically, conversational agents receive verbal input and produce verbal output in a conversational manner (Schöbel et al., 2023). While chatbots are characterized by their text-interfaces, social robots distinguish themselves by the physical embodiment (Chen et al., 2020). On the other hand, digital avatars are typified by their virtual embodiment, whereas virtual assistants, such as Siri and Alexa, are typically not visually represented (Adamopoulou & Moussiades, 2020). While most agents use text-based communication, others communicate using audible speech and visual modalities, such as gesticulation and facial expressions (ter Stal et al., 2020). Yet perhaps the most critical feature for distinguishing conversational agents is the dialogue software enabling conversations with human users. Rule-based agents operate using preprogrammed dialogue options and pattern-matching techniques to select suitable responses to human prompts (Schöbel et al., 2023). More advanced conversational agents are based on artificial intelligence (AI) technologies or, more specifically, large language models (LLM), which are driven by machine learning and autoregressive techniques (Adamopoulou & Moussiades, 2020; Allouch et al., 2021). Unlike rule-based systems, LLM-based agents are not constrained by pre-programmed responses, but generate their responses based on probabilistic techniques, providing more dynamic, individually tailored, and natural-sounding conversations (Shahriar & Hayawi, 2023).

Although Eliza was launched in the 1960s, about 50 years passed before we began seeing studies of conversational agents in psychotherapy. A crucial milestone emerged when Fitzpatrick et al. (2017) published the findings of a randomized controlled trial that examined the effectiveness and feasibility of cognitive behavioral therapy (CBT) delivered by the rule-based chatbot *Woebot* (N = 70). In this study, the human therapist was replaced by the chatbot Woebot, which aimed to assess and use CBT techniques in a conversational

manner to facilitate more positive thinking styles. After two weeks of access to Woebot, the treatment group reported a significant decrease in symptoms of depression in comparison to receiving access to an e-book on depression (d = 0.44). Participants in the treatment group on average used Woebot 12 times (*range* = 8-18) throughout the period, indicating a consistent level of engagement with the chatbot. Thenceforth, numerous randomized controlled trials have been published and have demonstrated promising effectiveness (Fitzpatrick et al., 2017; Fulmer et al., 2018; He et al., 2023; Klos et al., 2021; Ly et al., 2017). Most recently, a meta-analysis of 32 randomized controlled trials found significant improvements from mental health interventions delivered by conversational agents in symptoms of depression (k = 8, g = 0.48) and symptoms of generalized anxiety (k = 6, g = 0.35) compared to measurements only (He et al., 2023).

However, interpreting these meta-analytical findings is challenging. First, the meta-analysis included a heterogeneous variety of studies, distinguishable in terms of population, treatment approach and duration, primary outcome, and more (He et al., 2023). This mixture makes direct comparisons to meta-analyses with narrower scopes difficult, for example meta-analyses from psychotherapy research. Second, the reasons for the observed outcomes are uncertain, as the active ingredients in such treatments remain to be explored. As a consequence, there is limited basis for interpreting these findings and improving CA-psychotherapy, for example, in terms of what therapeutic strategies and techniques are more or less important to incorporate into conversational agents. The present contribution aims to help establish such a process-oriented research direction.

2.2. The active ingredients in psychotherapy

Most studies in this field have examined traditional CBT delivered by conversational agents (He et al., 2023). This trend likely reflects its recognition as the 'gold standard' of therapies (David et al., 2018) and the convenience of implementing its manual-based format in digital technologies. Despite this status, however, researchers should not take for granted that traditional CBT – or other specific therapeutic approaches – work exactly as theoretically prescribed. Indeed, there are substantial controversies surrounding what elements of psychotherapy make it effective (Cuijpers et al., 2019; Grencavage & Norcross, 1990; Kazdin, 2007; Wampold & Imel, 2015). The present section outlines prominent perspectives on the active ingredients in psychotherapy to construct an analytical framework for the narrative review. Our overarching claim here is the following: The controversies surrounding the active ingredients in psychotherapy necessitate an analytical framework that acknowledges multiple, sometimes conflicting theoretical perspectives as explanatory frameworks for CA-psychotherapy. On this basis, we derived search terms from a wide array of theoretical perspectives for our systematic literature search.

The main therapeutic approaches in psychotherapy are the so-called *bona-fide* approaches, such as traditional CBT, psychodynamic therapy, and interpersonal therapy (Braun et al., 2013; Wampold et al., 1997). Bona fide therapies are commonly considered 'validated' treatments as their efficiency has been established in numerous studies and they are based on a 'clear' rationale (Braun et al., 2013). As briefly mentioned earlier, the elements in psychotherapy that foster therapeutic change (e.g., symptom reduction) are called the *active ingredients* (Cuijpers et al., 2019; Lorenzo-Luaces, 2023). A common example is cognitive restructuring techniques, where the therapist, for instance, challenge the validity of the client's beliefs to foster more positive thinking (Powers et al., 2017). The psychological change processes elicited from such therapeutic techniques are referred to as *mechanisms of change* (Kazdin, 2007), for example less dysfunctional beliefs about oneself and the world. In process-research, the active ingredients believed to characterize specific therapeutic approaches are the so-called *specific factors* (Cuijpers et al., 2019; McAleavey & Castonguay, 2015). Theoretically, specific factors are designed to relieve an etiological *cause* of the disorder (Wampold & Imel, 2015). For example, an assumption in traditional CBT is that dysfunctional beliefs elicit emotional issues, and cognitive restructuring techniques are therefore employed to alter these beliefs (Powers et al., 2017). The implicit assumption that an etiological cause must be relieved to experience improvement is sometimes referred to as the *medical model* due to its resemblance to medical theories of pathology and treatment (Elkins, 2009; Wampold & Imel, 2015). An implication of this assumption is that, in theory, each therapeutic approach fosters therapeutic change through different causal pathways (Wampold & Imel, 2015). However, this assumption has been contested both theoretically and empirically (Cuijpers et al., 2017; Frank & Frank, 1991; Grencavage & Norcross, 1990; Lemmens et al., 2017; Lemmens et al., 2016; Wampold & Imel, 2015).

Despite different etiological assumptions and treatment strategies across approaches, however, identifying differential effectiveness and unravelling the active ingredients have been challenging (Cuijpers et al., 2019). Even though CBT is commonly considered the gold standard, meta-analyses have found none to small differential effectiveness across bona fide therapies (Cuijpers, 2017; Wampold et al., 1997), especially when controlling for researcher allegiance (Luborsky et al., 2002). Rosenzweig (1936) observed this trend and coined it the dodo bird verdict as an analogy to Alice's Adventures in Wonderland in which a dodo bird concluded a race by announcing "*Everyone has won, and all must have prizes!*". In other words, the dodo bird verdict expresses the empirical trend that, despite conceptual differences, bona fide therapies demonstrate comparable efficiency (Cuijpers et al., 2019).

Partly owing to the dodo bird verdict, numerous scientists have contested the idea of distinct causal pathways across therapeutic approaches. Broadly speaking, the so-called *common factors movement* claims that therapy works through shared rather than distinctive

causal pathways (Cuijpers et al., 2019; Wampold & Imel, 2015). Common factor approaches emphasize the common active ingredients across approaches, also called the *common factors*, such as therapeutic relationship and expectations of improvement (Wampold & Imel, 2015). Common factors arise not because of the deliberate intention of the therapist's technique but due to similarities in the structure, procedures, and type of social interaction that are shared across therapeutic approaches (Garfield, 1995). As an illustration, a close, trusting relationship to a warm, empathetic therapist is considered to be helpful in several common factor approaches (Wampold & Imel, 2015). Some approaches also highlight the benefits of positive outcome expectations arising partly from the sociocultural context surrounding therapy (Frank & Frank, 1991; Wampold & Imel, 2015). In other words, common factors are not bound to the therapist's techniques but rather inherent features of the therapeutic setting regardless of therapeutic orientation, whether that be CBT, psychodynamic therapy, or some other approach.

Yet it is important to note that the common factors movement does not represent a unitary front but comprises a wide array of scientists with more or less distinctive assumptions. Notably, Grencavage and Norcross (1990) conducted a review of 50 publications of different authors and revealed 89 common factors conceptualized in the literature. The most frequently proposed common factor was the interpersonal relationship (k = 26). Other frequent common factors were emotional ventilation, i.e., getting relief from negative emotions by expressing them (k = 19) and expectations/hope (k = 13). On the other hand, there was considerably less consensus regarding techniques (k = 7) and adherence to a treatment protocol (k = 4). Although consensus has limited epistemic value, these findings illustrate the abundance of common factors proposed in the literature. Among common factors approaches, however, perhaps the most influential framework nowadays is the *contextual model*. In this approach, Wampold and Imel (2015) claim that the effectiveness of therapy emerges from the *real* relationship (i.e., a genuine relationship where each perceives each other accurately) with a therapist, positive outcome expectations, as well as the specific factors of each therapy. However, Wampold and Imel (2015) did not consider therapeutic techniques effective because they address some hypothesized cause of the disorder. Rather, because the techniques are beneficial regardless of how the condition came about. For example, thinking more positively about oneself may be beneficial regardless of whether dysfunctional self-referential beliefs originally caused one's emotional problems.

A core issue feeding into the controversies surrounding psychotherapy is problematic research methods and ambiguous research findings. Experimental manipulations of specific factors (i.e., component studies) tend to show null findings, suggesting limited to none isolated effect of specific factors such as cognitive restructuring (Bell et al., 2013; Cuijpers et al., 2017). While insufficient statistical power is a possible alternative explanation of these

null findings (Cuijpers et al., 2017), the external validity of component studies is also limited, as treatment components might interact with each other rather than exerting isolated influences on the outcome. For example, cognitive restructuring exercises may be more effective when coupled with exposure exercises, as the latter may provide 'evidence' that the dysfunctional assumptions are unfounded. Another common process-research strategy is mediation analyses. Recent years have featured an increasing number of time-lagged mediation analyses to ensure temporal distance between the variables of interest (Hayes et al., 2022; Johannsen et al., 2022a), thereby strengthening interpretations of causality. While time-lagged mediation studies exhibit a relatively high epistemic value, they yield mostly weak or no supported for the causal pathways of therapeutic approaches. For example, CBT (A-Tjak et al., 2021; Lemmens et al., 2017; Quigley et al., 2019), acceptance and commitment therapy (A-Tjak et al., 2021; Johannsen et al., 2022b), interpersonal therapy (Lemmens et al., 2017; Lemmens et al., 2016), and mindfulness-based cognitive therapy (Johannsen et al., 2022b) for depression or anxiety have provided little to no time-lagged evidence in support of the hypothesized mechanisms of action. Moreover, although the alliance emerges as one of the strongest predictors of treatment outcomes (d = 0.57 (Horvath et al., 2011; Wampold & Imel, 2015)), a meta-analysis of time-lagged mediation research also shows mixed results for the alliance (Baier et al., 2020). Such findings are in conjunction with the uncontroversial notion in process-research that there is limited firm, empirically substantiated evidence for how psychotherapy works (Cuijpers et al., 2019; Kazdin, 2007; Lemmens et al., 2016; Lorenzo-Luaces, 2023; Mulder et al., 2017; Wampold & Imel, 2015). As a consequence, a cautious approach is necessary when constructing an analytical framework for the present narrative review.

2.3. The present narrative review

Although the field of CA-psychotherapy is currently flourishing, our understanding of the active ingredients remains scarce. On this basis, this narrative review aimed to delineate current evidence for the active ingredients in CA-psychotherapy and provide directions for future research. Given the controversies connected to establishing the active ingredients, we acknowledge conflicting perspectives (i.e., the medical model, such as CBT, and common factor approaches) as explanatory frameworks for the effectiveness of psychotherapy. On this basis, the current narrative review exploratively evaluated the evidence for the active ingredients in CA-psychotherapy. To this end, we derived search terms from a wide array of theories of psychotherapy (see **Appendix**). We also included studies that investigated human-computer interactions in contexts beyond psychotherapy. Our main reason for this decision was to assess whether the preconditions exist for applying the theoretical frameworks from psychotherapy in this field. For example, research showing that individuals

tend to self-disclose to conversational agents could provide indications of emotional venting in CA-psychotherapy.

3. Method

We conducted a narrative review of the scientific literature, as this methodology allowed us to exploratively identify the emerging research trends in this field. Specifically, we conducted a systematic literature search on four electronic bibliographic databases (PsychInfo, Pubmed, ACM, and IEEE) up until May 15, 2023. The search strategy was employed for the four databases independently using Boolean operators. To ensure comprehensiveness, the search terms were derived from reviews and theoretical papers (Grencavage & Norcross, 1990; Hayes et al., 2006; Johannsen et al., 2022b; Kirsch et al., 2016; Powers et al., 2017; Røssberg et al., 2021; Wampold, 2015; Wampold & Imel, 2015) and discussed within the research team. Two researchers in psychology and psychotherapy not authoring the present paper provided suggestions for other terms to include in the search string. To minimize the number of irrelevant results, terms with multiple definitions and usages (e.g., 'acceptance') were intentionally omitted from the search. The final search string included different terms for 'conversational agents' (e.g., chatbots and virtual avatars) and active ingredients as suggested in various theoretical approaches (see **Appendix** for a full overview of search terms). As can be seen from **Appendix,** the search terms contained specific terms for 'active ingredients' and 'mechanisms of change', as well as specific examples of hypothesized factors in psychotherapy (e.g., cognitive restructuring, decentering, alliance, etc.). This approach was chosen because more general terms, such as 'active ingredient', are not consistently used in the human-computer interaction literature. Furthermore, this approach was also chosen to include studies on contexts beyond CA-psychotherapy, as previously argued. The search was conducted on titles, abstracts, and keywords. We limited the search to records published within the last 10 years (from May 2013) to ensure that we excluded studies involving (severely) outdated technology. Snowballing and Google searches was used to search for additional relevant records.

All references were imported to the review management software *Covidence systematic review software, Veritas Health Innovation, Melbourne, Australia* (for flowchart see **figure 1**). Two reviewers (MASKED) independently screened all titles and abstracts in accordance with the inclusion-and exclusion criteria outlined below. Following the abstract and title screening, the remaining records were independently screened in full text by the same two authors. Disagreements on eligibility were resolved by discussion. Study selection was based on predetermined inclusion and exclusion criteria. To be eligible for inclusion, studies had to meet the following inclusion criteria: 1) peer-reviewed; 2) English language; 3) experimental design; 4) experimental manipulation *or* measurement of a hypothetical unique, common factor or mechanism of change in psychotherapy; 5) a direct interaction

between human participants and a conversational agent, which was defined as a software system that mimics a natural verbal conversation with human users (i.e., agents that respond to and generate verbal language).

Studies were excluded if: a) more than 10% of the study population consisted of children (0-17 years) or older adults (65+ years). The 10% threshold was set to prevent the exclusion of studies where the vast majority of participants fell within the target age group, while also minimizing the potential for developmental variations to confound the evidence base; b) the study population was defined by the presence of somatic illness, neurodevelopmental disorders (e.g., autism spectrum disorder, attention-deficit/hyper-attentive disorder, intellectual disability), severe mental illness (i.e., bipolar disorder, schizophrenia, eating disorder), or substance use disorder. This criteria was chosen to confine the narrative review to emotional conditions, such as depression, anxiety, and psychological distress, as promising meta-analytical evidence has emerged for such conditions among adults (He et al., 2023); c) the primary outcome was health behavior (e.g., eating habits, vaccine resistance, exercise, smoking, etc.); d) the conversational agent was remote-controlled during the intervention. This criterion was included to increase generalizability of the selected studies to current CA-psychotherapy platforms. Reasons for exclusion were noted for each excluded record during full text screening.

4. Findings and discussion

[TABLE 1]

The systematic literature search in the four databases yielded a total of 911 publications (see **Figure 1** below). A Google search yielded an additional record. After removal of duplicates, 835 records remained. In the abstract screening, 766 records were excluded (*Kappa* = .71). In the full-text screening, 52 records were excluded (*Kappa* = .65)[1]. In total, 17 records met the inclusion and exclusion criteria. See **Table 1** for an overview of the included records. See **Appendix** for a full overview of selected records, including methodological characteristics and outcomes statistics.

[FIGURE 1]

Across the 17 studies, three overarching themes emerged: relationship variables, emotional venting, and cognitive mechanisms. Common factor approaches as well as specific therapeutic theories (e.g., CBT) functioned as the analytical framework for delineating common themes in the literature, as argued previously. Based on those theoretical approaches, the themes were composed by evaluating the conceptual similarities across the studies. The theme *relationship variables* emerged from selected studies that investigated alliance or empathy, as the latter has been suggested to antecede the alliance (Wampold &

[1] Kappa coefficients can be interpreted as follows: ≤0 = poor, .01–.20 = slight, .21–.40 = fair, .41–.60 = moderate, .61–.80 = substantial, and .81–1 = almost perfect (Landis & Koch,1977).

Imel, 2015). The theme *emotional venting* was derived from studies on self-disclosures to conversational agents. The theme *cognitive mechanisms* emerged from a single study that examined a therapy chatbot utilizing cognitive restructuring or cognitive defusion techniques to alter dysfunctional thinking. See **Table 2** for a brief overview of the main findings summarized as bullet points.

[TABLE 2]

4.1. Relationship variables

4.1.1. Theory

Numerous conceptualizations of the relationship between therapist and client exist (Horvath et al., 2011). Perhaps the most influential definition was coined by Bordin (1979), according to whom the so-called alliance denotes three interrelated components: agreement on therapy goals, agreement on therapy tasks, and the bond between therapist and client. The establishment of a strong bond between therapist and client has been suggested to rely on the therapist's manifestation of empathy *(Wampold & Imel, 2015)*, which involves the ability to perceive and share the feelings of others (Rogers, 1951; Watson, 2016). Barrett-Lennard (1993) proposed several stages of empathy, ranging from the therapist sensing and communicating their understanding of the client's feelings to the client perceiving the empathetic response and experiencing a feeling of being understood and validated. In the contextual model, the therapist's expression of empathy is considered necessary to establish a strong emotional bond (Wampold & Imel, 2015).

Besides increasing client adherence and engagement, a strong relationship between therapist and client may also influence the treatment outcome through causal pathways. One possible pathway is that the emotional bond, sometimes conceptualized as a 'real relationship', between therapist and client might contribute to fulfilling basic human social needs (Wampold & Imel, 2015). It is widely acknowledged that humans have a basic psychological need for experiencing a sense of connectedness to others, such as attachment (Bowlby, 1980), relatedness (Deci & Ryan, 2000), or perceived social support (Wills & Shinar, 2000). Relatedness is contrasted by loneliness, which has been shown to be a robust predictor of mental health (Wang et al., 2020). As highlighted by Wampold and Imel (2015), the emotional bond to the therapist might particularly help individuals who lack close relationships.

A second possible pathway is corrective emotional experiences. Based on an etiological assumption in psychoanalysis that chaotic interhuman relationships from the past can cause dysfunctional beliefs about oneself (e.g., '*I am not worth caring for*') and about being in relationships to others (e.g., '*Others cannot be trusted*'), the supportive and trusting relationship to a therapist is believed to challenge the dysfunctional beliefs (Jørgensen, 2004). The clients experience that they can express their needs and feelings in close

relationships without being rejected, which might be particularly beneficial to individuals with a history of chaotic relationships and an insecure attachment style (Bohart & Greenberg, 1997).

Finally, the physical presence of another person that responds in an empathetic, warm manner may itself reduce emotional distress. Emotional co-regulation is the dyadic phenomena where heightened emotional arousal returns to a baseline level due to the care and comfort provided by another person (Soma et al., 2020; Wampold, 2021). Although the positive implications of emotional co-regulation could be short-lived, it might facilitate a trusting emotional bond, as well as corrective emotional experiences.

4.1.2. Findings

We identified seven studies that explored relationship variables (see **Table 1** for a full overview of study characteristics), including four randomized controlled trials (de Gennaro et al., 2020; He et al., 2022; Johanson et al., 2020; Liu et al., 2022), two pretest-posttest experimental studies (Ellis-Brush, 2021; Jeong et al., 2020), and one between-condition experiment for which the allocation procedure was not reported (Al Farisi et al., 2022). All studies tested chatbots with rule-based dialogue software.

Four studies examined the alliance with a therapy chatbot (Ellis-Brush, 2021; He et al., 2022; Jeong et al., 2020; Liu et al., 2022). Jeong et al. (2020) examined the alliance to a social robot, but they did not explore the relationship between the alliance and outcome. Ellis-Brush (2021) tested the alliance to the Wysa chatbot, yet neither reported inferential analyses of the relationship to therapy outcomes. Liu et al. (2022) found that the alliance to the therapy chatbot XiaoNan was significantly better than to an e-book. He et al. (2022) showed that the alliance to the therapy chatbot XiaoE was significantly better than to a general-purpose chatbot.

Furthermore, two studies examined antecedents of perceptions of empathy in chatbots. Johanson et al. (2020) assessed the role of humor in chatbots, but they did not report statistical analyses of between-condition empathy levels (see **Table 1**). Al Farisi et al. (2022) found that an empathetic chatbot demonstrating anthropomorphic behaviors was rated as significantly more empathetic than a neutral chatbot. The final study looked into the emotional implications of empathetic behaviors in a chatbot. de Gennaro et al. (2020) showed that interacting with an empathetic chatbot significantly improved mood following a social exclusion experience as compared to self-disclosing in an interactive questionnaire.

3.1.3. Discussion

Overall, there was no compelling evidence concerning what role(s) the (potential) human-computer relationship plays in CA-psychotherapy. Particularly, the role of the alliance remains tentative as none of the included studies statistically explored the relationship between the alliance and therapy outcomes. As such, there is no empirical

evidence to suggest the alliance plays any role in CA-psychotherapy. On the other hand, there was evidence to suggest that empathetic responses in chatbots improve mood after experiencing a decrease in mood, signaling a possible role of empathy in CA-psychotherapy. However, the lack of a matching comparison condition (e.g., that involved a chatbot interaction) suggests that the novelty and excitement, also called the *novelty effect* (Smedegaard, 2022), potentially linked to an initial interaction with a chatbot is a possible confounding variable.

3.1.4. Implicit assumptions of relational processes in psychotherapy

While specific methodological issues challenge the idea that the relationship somehow serves a healing function in CA-psychotherapy, there are also implicit assumptions in psychotherapy that cannot be taken for granted when investigating CA-psychotherapy. Core questions remain unanswered, for which the answers can have significant implications for whether conversational agents can assume the roles of human therapists in all respects.

The first question is whether humans can form affectionate bonds to conversational agents, which resemble interhuman relationships. When studying psychotherapy, researchers can take for granted that humans can form affectionate bonds with each other, and that such bonds can have positive psychological implications. These are premises for assuming that an alliance (as conceptualized in psychotherapy theories) can be established in therapy, and that the alliance somehow promotes therapy outcomes. On the other hand, although conversational agents can reenact human social behaviors (e.g., verbal communication, turn-taking, and empathetic actions), it remains to be determined if humans possess the capabilities for forming affectionate bonds to machines in a manner that resemble interhuman relationships. While qualitative research has shown that some individuals describe signs of strong relationships to conversational agents (Loveys et al., 2022; Pentina et al., 2023; Skjuve et al., 2021, 2022), other research found evidence to suggest that this attachment does not resemble interhuman relationships. Indeed, Pentina et al. (2023) found that individuals who of their own accord sought companionship with Replika experienced a stronger attachment to it yet felt lonelier than individuals who had been instructed to use Replika for two weeks. Two non-mutually exclusive explanations may account for this trend: 1) loneliness is an antecedent of seeking companionship in Replika *and* 2) companionship with Replika does not alleviate loneliness. Another potential explanation is that unobserved individual-level variables distinguishing the conditions accounted for the observed difference in loneliness, signaling the need for tighter controlled experiments on this subject.

Nevertheless, there could be inherent characteristics of conversational agents that render them unable to resemble interhuman relationships. As noted by Wieland (2023), conversational agents may not qualify as relational partners, as there is limited reciprocity in

human-computer relationships. In her point of view, for an agent to be considered a relational partner, there must be a reciprocal exchange of care and empathy, and the absence of selfhood in conversational agents implies that there is nothing to care about for the human counterpart (Wieland, 2023). Likewise, the formation of a 'real relationship', as conceptualized by Gelso (2014), necessitates not only that the counterpart acts as a human but also that the friendliness and care are perceived as *genuine*, i.e., the friendliness and care reflects authentic positive sentiments. When there is no 'other', however, there is nothing to perceive as genuine, implying that the conceptual criteria for a 'real relationship' are not fulfilled. As a consequence, the expressed friendliness and care may not possess the same value, just as the inauthentic friendliness expressed by another person may be valued less than perceived authentic friendliness. Such theoretical possibilities imply that researchers should remain skeptical about whether the alliance can serve similar functions in CA-psychotherapy as in traditional psychotherapy.

The second question is whether humans acquire social learning experiences from human-computer interactions, which can be transferred to interhuman relationships. If not, it seems unlikely that CA-psychotherapy can foster corrective emotional experiences. As mentioned, a corrective emotional experience is a form of social learning experience, in which the client experiences that others can be trusted and will support them. Yet it is important to acknowledge the possibility that learning experiences might be bound to specific contexts. A related idea is the concept of situated cognition, which refers to the notion that knowledge is bound to situations in which they are acquired (Wilson, 1993). Hence, a core hypothesis in this line of thinking is that transfer of learning experiences across dissimilar situations is unlikely. One implication of this possibility is that it can be questioned whether social learning experiences acquired in human-computer relationships are transferable to interhuman relationships. From the client's point of view, a salient difference between humans and conversational agents could be that the latter does not possess consciousness. So, if the client's dysfunctional beliefs are rooted in relationships with conscious beings, engaging with an empathetic conversational agent may not challenge beliefs bound to interhuman relationships. However, it is important to note that the manifestation of anthropomorphic features may blur the differences between humans and conversational agents, phenomenologically speaking. The 'Computers as Social Actors Paradigm' anticipates that humans ascribe machines traits such as consciousness and empathy and apply scripts for social behaviors when interacting with anthropomorphic (i.e., human-like) agents (Gambino et al., 2020). In other words, when the conversational agent, phenomenologically speaking, resemble humans in some respects, one might forget that it is not a human. Hence, it is premature to reject the possibility that anthropomorphic

conversational agents demonstrating empathy and warmth, to some extent, can foster corrective emotional experiences.

4.2. Emotional venting

4.2.1. Theory

Emotional venting is the process of coping with emotions by expressing them, typically by describing them with words (Trần et al., 2023). This idea dates to the psychoanalytic tradition, in which it was believed to serve a healing function by discharging suppressed negative feelings (Jørgensen, 2004). Numerous researchers have suggested that emotional venting is a common factor in psychotherapy (Grencavage & Norcross, 1990). One possible reason is that describing one's thoughts and feelings can be considered a necessary step to change styles of thinking and feeling. It also seems plausible that expressing disturbing emotions can foster emotional coregulation through empathetic responses from the therapist. In contemporary human-computer research, the concept of self-disclosure has received much attention and to some extent represents the behavioral dimension of emotional venting. Self-disclosure refers to communicating information about oneself to others that is considered personal and/or sensitive by the disclosing individual (Ignatius & Kokkonen, 2007).

4.2.2. Findings

We identified nine randomized controlled trials that examined antecedents and/or implications of self-disclosing to conversational agents (see **Table 1**) (Akiyoshi et al., 2021; Kang & Kang, 2023; Lee et al., 2022; Meng & Dai, 2021; Pujiarti et al., 2022; Qian et al., 2019; Schuetzler et al., 2018; Yi-Chieh, Naomi, & Yun, 2020; Yi-Chieh, Naomi, et al., 2020). All studies tested rule-based conversational agents, some of which compared agents with different features (Kang & Kang, 2023; Pujiarti et al., 2022; Qian et al., 2019; Schuetzler et al., 2018; Yi-Chieh, Naomi, & Yun, 2020; Yi-Chieh, Naomi, et al., 2020).

Two studies examined the emotional implications of self-disclosing to a conversational agent. Akiyoshi et al. (2021) showed that self-disclosing to a chatbot significantly decreased anger, but did not influence other emotions, such as anxiety and depression. Meng and Dai (2021) found that self-disclosing to a human or chatbot (each of which providing emotional support and engaging in reciprocal self-disclosures) did not affect worry significantly differently.

Seven studies examined the antecedents of self-disclosing to conversational agents. Schuetzler et al. (2018) found that the participants were significantly more likely to report alcohol drinking behaviors to a chatbot than to a human. Chatbots asking relevant follow-up questions elicited more self-disclosures, and men were more likely than women to self-disclose to the chatbot. Qian et al. (2019) found no differences in self-disclosures from interacting with conversational agents featuring various communication modalities (e.g.,

speech vs. texting). In broad terms, (Yi-Chieh, Naomi, & Yun, 2020; Yi-Chieh, Naomi, et al., 2020) found that a chatbot highly engaging in reciprocal self-disclosures elicited more self-disclosures from participants and was rated higher in trust, enjoyment, and intimacy than non- and low-reciprocal self-disclosing chatbots. Akiyoshi et al. (2021) compared the number of self-disclosures to a chatbot utilizing the column method (i.e., a CBT method to help clients identify the nature and triggers of negative thoughts and their underlying self-schemes) versus a combination of the column method and additional conversational strategies, such as open-ended questions to promote reflection. The combined condition elicited significantly more self-disclosures. Pujiarti et al. (2022) found that conversational atmosphere visualizations and co-activity increased self-disclosures. Finally, Lee et al. (2022) found that perceived social presence of the agent and fear of negative evaluation predicted less self-disclosures to a chatbot.

4.2.3. Discussion

Overall, there is sparse evidence to suggest that self-disclosing to conversational agents has positive emotional implications. Only two studies examined the emotional implications of self-disclosing to conversational agents and primarily showed null findings (Akiyoshi et al., 2021; Meng & Dai, 2021). However, one study (categorized under the theme 'relationship variables') showed that self-disclosing to a chatbot that responds empathetically improves mood following a social exclusion experience. This indicates that not only the act of unburdening oneself but also empathetic and validating responses may account for the benefits of self-disclosures. Hence, considering the studies discussed here, there is mixed evidence to substantiate the idea that unburdening oneself to machines can be helpful. However, it is important to note that written self-disclosures, operationalized as expressive writing, is a well-established intervention for reducing distress (Smyth, 1998). Thus, coupled with the observed benefits of empathetic responses from chatbots (de Gennaro et al., 2020), it seems likely that self-disclosing to conversational agents programmed to respond empathetically can be especially helpful. Nevertheless, there is a pressing need to utilize stronger methodologies to examine the emotional implications of self-disclosing to conversational agents.

While few studies investigated the emotional outcomes of self-disclosures, several studies examined the antecedents of self-disclosures (Akiyoshi et al., 2021; Lee et al., 2022; Pujiarti et al., 2022; Qian et al., 2019; Schuetzler et al., 2018; Yi-Chieh, Naomi, et al., 2020). This line of research serves an important function in exploring what prompts people to unburden themselves from their personal issues to conversational agents, possibly highlighting the factors that facilitate the emotional implications. Notably, there was evidence to suggest that individuals disclose more sensitive information about themselves to chatbots than to humans (Schuetzler et al., 2018). This finding suggests that something

which phenomenologically differentiate conversational agents and humans influences our propensities to self-disclose to them. One possibility is the lay belief that conversational agents lack emotional capacities and thus the ability to form negative evaluations of others (Kim et al., 2022; Lucas et al., 2017). In line with this notion, research has demonstrated that *anthropomorphic features* (i.e., human-like features of the agent), as well as *anthropomorphism* (i.e., the attribution of human-like features), influence the propensities to self-disclose. Indeed, one of the selected studies found that chatbots represented by a humanoid embodiment elicited less honest self-disclosures than to disembodied chatbots (Kang & Kang, 2023). Another study found that perceived social presence and fear of negative social evaluation from chatbot negatively predicted self-disclosures to a chatbot (Lee et al., 2022).

Moreover, there may be additional fine-grained interactions accounting for the likelihood to self-disclose to conversational agents. Research (which did not meet our selection criteria) has shown that socially anxious individuals tend to disclose more information to a remotely controlled conversational agent than to humans, whereas no difference was found for non-socially anxious individuals (Kang & Gratch, 2010). Another possible antecedent is the motivation for self-disclosing. Kim et al. (2022) examined the role of motivation to avoid negative social evaluation and seek social support in relation to self-disclosing to AI agents and humans. They found that humans tend to self-disclose more to other humans than AI agents when seeking social support yet self-disclose more to AI agents when fearing negative social evaluations (Kim et al., 2022). A conversational agent may thus constitute a risk-free recipient to talk about one's sensitive issues and seek advice from when fearing negative social evaluations. However, conversational agents may not be considered suitable for seeking social support and validation. Taken together, it is crucial to acknowledge complex interactions between conversational agent features and individual differences and motivations. This underpins the necessity of investigating for whom and under what circumstances conversational agents can be helpful.

Besides the issues considered previously, numerous questions remain unanswered, such as the possible role of emotional bonding to conversational agents in terms of self-disclosures (Skjuve et al., 2022), and *how* self-disclosing to conversational agents might be beneficial. For example, whether self-disclosing serves an instrumental role through the responses one gets from the other or on its own is helpful has implications for how conversational agents should be programmed. Indeed, emotional venting as an emotion-focused extratherapeutic coping strategy has been demonstrated to be inversely related to mental health (Liverant et al., 2004). As a theoretical consequence, the high accessibility and on-demand access to conversational agents might pose a threat to mental health when individuals gain unlimited access to emotional venting. However, as suggested by Bohart

(1980), emotional venting can serve an instrumental role in promoting reflections on the triggers of disturbing feelings. When venting to others, there occurs an opportunity for others to challenge the beliefs responsible for the negative feelings, which, as predicted by the cognitive model (Beck, 1976), may serve a healing function. If this holds true, conversational agents should not only be designed to elicit self-disclosures and validate one's understanding of the world; they should also be designed to challenge the beliefs underlying the expressed feelings.

4.3. Cognitive factors

4.3.1. Theory

The theme 'cognitive factors' encompasses two overarching conceptual approaches across cognitive therapies. Theoretically speaking, cognitive therapies work by changing *what* people think (i.e., changing the *content* of thoughts) and *how* people think (i.e., changing the *function* of thoughts (Longmore & Worrell, 2007). To provide some brief illustrations, the effectiveness of traditional CBT is, as mentioned, believed to be driven by changing what people think about themselves and the world. To foster this change, cognitive restructuring techniques are employed (Fenn & Byrne, 2013). On the other hand, therapists delivering acceptance and commitment therapy, amongst other things, aim to change the function of thoughts by fostering a neutral, observing stance towards one's thoughts rather than becoming entangled, or 'fused', with them (i.e., cognitive defusion techniques (Hayes et al., 2006).

4.3.2. Findings

We identified one randomized controlled trial that examined cognitive factors in therapeutic interventions delivered by a rule-based chatbot (See **Table 1**). Lavelle et al. (2022) experimentally manipulated cognitive restructuring and cognitive defusion techniques, as well as employed standardized tools to assess mood and variables indicative of cognitive style. However, no significant increases were observed at follow-up (Lavelle et al., 2022).

4.3.3. Discussion

There are at least two explanations for these null findings. First, because of drop-out and exclusion, only 28 responses out of 223 were analyzed, which increases the risk of Type II errors, for instance, due to low statistical power and heightened susceptibility to random variability. We conducted a sensitivity power analysis and found that the study was not powered to detect statistically significant two-tailed differences below $d = 1.33$ ($\alpha = .05$, $1\text{-}\beta = .80$, $n_1 = 9$, $n_2 = 11$). Second, the study population was students without mental disorders. This implies that the targeted assumptions of the clients might not have been sufficiently dysfunctional for the cognitive restructuring techniques to make an impact. Likewise, it seems plausible that the efficiency of cognitive defusion techniques depends on the client's

tendencies to become entangled with negative thoughts. Given these circumstances, this study does not challenge the possibility that cognitive factors are effective in CA-psychotherapy, nor does it support it.

4.4. Limitations of the narrative review

We would like to acknowledge some limitations of this narrative review. The first limitation concerns our choice of including studies in contexts beyond therapy.
One core implication of this choice is that the findings cannot necessarily be directly generalized to CA-psychotherapy. The reason why is that several of the included studies examined micro-interactions of a few minutes' duration. Therefore, those studies may not inform us about how individuals engage with conversational agents longitudinally. To illustrate why, qualitative research suggests that individuals form stronger relationships with conversational agents over time (Skjuve et al., 2022), particularly from the initial interactions and forth, which might enhance engagement and thus the effectiveness of therapeutic techniques. As such, in the absence of more studies taking place in a multi-session psychotherapeutic setting, this evidence base remains weak due to limited generalizability to multi-session interventions.

Second, some of the included studies had no mental health outcome variable. This restricts firm interpretations of the role of the supposed active ingredient. We refrained from limiting the study selection to studies with mental health outcome variables to avoid excluding potential studies that examined variables indicative of mechanisms of change, such as changes in dysfunctional thoughts. This strategy ensured a comprehensive literature overview yet entailed that some of the selected studies provide limited insight into the active ingredients in CA-psychotherapy. Nevertheless, these studies still yield valuable epistemic insights relevant to this review. Indeed, they highlight, for example, that the tendency to self-disclose to conversational agents is determined by several circumstances, thereby potentially showcasing the predictors of emotional venting in CA-psychotherapy.

Third, the assessments made in this review are to some extent based on outdated technology. We find ourselves in a time where technology develops at an extreme pace. In November 2022, ChatGPT was launched, which demonstrated the impressive conversational capabilities that can be achieved with LLMs. Since no studies of LLM-agents were published at the time of our literature search, all studies considered in this review tested rule-based agents that lack the sophisticated conversational abilities of LLM-agents. As such, this review represents the best available knowledge, yet it is quite possible that this research field must be reevaluated in a few years.

The final limitation concerns the question of mental health interventions delivered by conversational agents should be understood as psychotherapy or something else. This is an important question, as a core assumption of this narrative review is that psychotherapeutic

theory, at least to some extent, can inform us about the active ingredients in interventions delivered by conversational agents. However, if this assumption turns out to be problematic, it could be misleading to discuss whether conversational agent-delivered psychotherapy can work through relational pathways. Our reason for making this objection is that psychotherapy has been highlighted as an *interpersonal* activity, distinguishing psychotherapy from interventions such as bibliotherapy and mindfulness meditation (Frank & Frank, 1991; Norcross, 2002; Wampold & Imel, 2015). In this perspective, it becomes necessary to consider conversational agents as *individuals*, at least from the clients' perspective, to justify applying psychotherapeutic theories. Several researchers have touched upon this issue. Sedlakova and Trachsel (2022) highlighted the hybrid nature of conversational agents in that they possess features as both agents and tools, lying somewhere in between those poles. They claim that conversational agents cannot promote therapeutic change as new insights are fostered by talking with a therapist that possesses mental capacities, emotions, and empathy. In other words, they highlight the *inherent* abilities of an agent as crucial in promoting therapeutic change. Hurley et al. (2023) challenged this position by arguing that the *phenomenological experience* of the conversational agent is more important than the technical processes leading to the demonstration of agency. They claim that the information the client receives is instrumental in fostering therapeutic change, and that it is less important whether a human or conversational agent is the sender of this information (Hurley et al., 2023). However, perhaps does the phenomenological experience of something not only rely on what it does, and how it appears, but also *what it is*. As highlighted by Kim et al. (2022), humans may have beliefs of what conversational agents are suitable for based on their lack of emotional capacities and inability to make interpersonal judgments. Perhaps due to such reasons, they found that individuals are less inclined to share sensitive information to a chatbot when seeking social support (Kim et al., 2022). Following this idea, lay beliefs about conversational agents may influence what meanings humans ascribe to their actions. Although conversational agents can be programmed to perform emotional support, this support might be perceived differently due to the belief that it stems from algorithms rather than genuine empathy. In other words, because there is nothing within conversational agents to hold positive sentiments. For these reasons, the question of whether psychotherapy is a suitable theoretical framework for this subject remains open.

5. Future directions

Our knowledge of the active ingredients in CA-psychotherapy is tremendously sparse. There is a long way to go before this field reaches the same state as process-oriented psychotherapy research. This section aims to help establish the foundation for advancing our knowledge of the active ingredients in CA-psychotherapy. To that end, we outline a series of

recommendations for future research. We acknowledge that some of these recommendations are somewhat generic, but they are borne out of a wish to aid future research in avoiding some of the methodological issues observed in this narrative review. Improved methodological awareness along these lines can also assist future research in effectively utilizing the potentials of conversational agents to enhance its epistemic value. As such, several of these recommendations are also relevant to research dedicated to the effectiveness of CA-psychotherapy.

[TABLE 3]

Recommendation 1: Minimize threats to internal validity in process-research

The first recommendation concerns the internal validity of process-oriented research in psychotherapy. Given the studies considered in this review, more awareness should be given to the requirements for establishing causality. Especially the studies investigating the alliance could have benefitted from more awareness of this (Ellis-Brush, 2021; He et al., 2022; Jeong et al., 2020; Liu et al., 2022). A highly cited set of recommendations for increasing internal validity in process-oriented psychotherapy research was outlined by Kazdin (2007). Essentially, Kazdin (2007) put forth a range of requirements for establishing causal relationships between mechanisms of action and treatment outcomes in psychotherapy. Although these criteria are not flawless, adhering to them can help disentangle how CA-psychotherapy brings about therapeutic change.

Kazdin (2007) suggested there should be a combination of the following circumstances within a line of research to establish a mechanism of change: 1) a 'strong' (whatever that is) association between the intervention, mediator, and outcome; 2) specificity of this association, such that only the hypothesized mechanism, or mediator, relates to the outcome and not a multitude of mediators (although this requirement may be difficult, if not impossible, to satisfy with correlational designs); 3) consistent findings across studies utilizing different samples and conditions. Researchers should though be aware that, theoretically speaking, there may be cases of moderated causation, where a mechanism only operates in a specific client population due to a certain etiology (Wampold & Imel, 2015); 4) experimental manipulation of the mechanism, for instance by utilizing the component study design. Here it is paramount to maximize the similarities between conditions so that only the variable of interest distinguishes them. Yet it should be noted, as discussed previously, that fragmentating interventions pose a threat to external validity; 5) a time-lagged relationship between cause and effect to rule out reverse causation (though not epiphenomena); 6) a gradient, that is, a dose-response relationship, although, as Kazdin (2007) note, the possibility of non-linear relationships should be taken into account; 7) plausibility, referring to the notion that the causal relationship must be theoretically sound, which, for instance,

rules out approaches based on pseudoscientific theories. Researchers should though be aware that this requirement clashes with what has been coined the demarcation problem, signifying the issue of determining the criterions for distinguishing between scientific and pseudoscientific theories (Lilienfeld et al., 2015). Nevertheless, a reasonable rule of thumb is that bona fide modalities can be considered theoretically sound approaches. Whereas time-lagged mediation research provides indications of mechanisms of change, component studies help delineate active ingredients.

Recommendation 2: Utilize the accessibility of digital conversational agents to increase statistical power

As previously discussed, insufficient statistical power might explain several controversial findings in psychotherapy research, including *the dodo bird verdict* and the sparse support for active ingredients in component studies (Cuijpers et al., 2017; Cuijpers et al., 2019). This issue arises from the large resources required to conduct multi-session psychotherapeutic interventions administered by human therapists. In contrast, CA-psychotherapy constitutes a low-cost, in-home, on-demand treatment platform, creating opportunities to conduct large-scale studies that can detect even subtle effects. Further underpinning this opportunity, the dialogues of (rule-based) conversational agents can be standardized to a greater extent than in psychotherapy, which increases control of confounding variables. As such, research on CA-psychotherapy could also provide insights into the active ingredients in psychotherapy.

Recommendation 3: Be aware of the multifaceted nature of conversational agents

The term 'conversational agent' is a multidimensional one, distinguishable in terms of visual representation, communication modality, dialogue software, and more. Researchers should therefore be cautious when trying to generalize findings across conversational agents. For example, the present narrative review identified several studies that showed differential effects of various features of conversational agents, such as anthropomorphic features (e.g., embodiment), displayed gender and personality (Kang & Kang, 2023), conversational style (Akiyoshi et al., 2021; Schuetzler et al., 2018), as well as features only applicable to conversational agents, such as conversational atmosphere visualizations (Pujiarti et al., 2022).

Besides this, conversational sophistication should also be taken into account. While current research primarily relies on rule-based conversational agents, technological advancements in recent years offer more sophisticated, LLM-agents, for instance based on the GPT language models (Brown et al., 2020). AI technologies are right now revolutionizing human-computer interactions, which is why we should be cautious of generalizing existing

evidence across agents. Yet it is crucial to also recognize differences within the categories of rule-based and AI-based agents. Indeed, rule-based agents are not per se inflexible; their dialogue flows can range from highly complex to poorly designed, where few response options are scripted and limited *a priori* testing has been conducted.

Hence, as the term 'conversational agent' is not a unitary one, it is challenging to generalize from the specific to general levels of abstraction. This conceptual diversity further complicates the discussion of whether and how conversational agents are effective in the context of therapy.

Recommendation 4: Take the diversity of human reference groups into account

Some studies identified in our literature search utilized human comparisons to assess the potential of conversational agents (Meng & Dai, 2021; Schuetzler et al., 2018). While such studies provide insights into whether conversational agents, broadly speaking, are more or less effective than human therapists in some respects (e.g., eliciting self-disclosures), it is important to recognize individual differences as a confounding variable. Research has shown significant *therapist effects*, denoting the variability across therapists with respect to the efficacy of treatments (Saxon et al., 2017). For example, the therapists' interpersonal skills have been suggested to play a key role to therapy outcomes (Norcross, 2019). While an early meta-analysis showed that the therapist accounted for 8% of therapeutic outcomes (Crits-Christoph et al., 1991), a more recent meta-analysis found it to predict 5% (Baldwin & Imel, 2013), the latter corresponding to an effect size of $d = 0.46$. When including human therapist as comparison conditions in evaluations of the human-computer relational processes, it is therefore important to be aware of the individual differences across therapists. This recommendation also concerns individual differences in eliciting self-disclosures, such as the abilities to create a trusting environment.

Recommendation 5: Assess clinical populations

Multiple studies considered in this review did not examine active ingredients in a psychotherapeutic setting. Of those studies that did, several enrolled individuals without mental health issues (Ellis-Brush, 2021; Jeong et al., 2020; Lavelle et al., 2022), which poses challenges for generalizing the results to those populations that receive therapy. Consistent with the medical model, the effectiveness of psychotherapy might be determined by the presence of a psychological problem (e.g., dysfunctional assumptions) that the techniques are believed to remedy. Therefore, findings from studies including non-clinical populations can give an erroneous impression of the evidence for an active ingredient and the effectiveness of a treatment. The external validity findings of studies including healthy, non-clinical populations can therefore be questioned, which is why we recommend that

researchers aim at maximizing the similarities between the study population and intended target population. Of course, there can be ethical challenges associated with studying clinical populations due to their emotional vulnerability coupled with the possibility of adverse effects yet to be discovered. At some point, however, it becomes necessary to study clinical populations to advance our knowledge of the utility of conversational agents.

Recommendation 6: Preregister your studies to increase the credibility of our evidence base

The limited number of preregistered studies (k = 2) identified in the literature search raises concerns about the credibility of this evidence base. One risk associated with non-preregistered studies is p-hacking, where the researchers selectively choose to report those analyses that yield statistical results (Head et al., 2015). Another risk when there is no tradition for preregistration within a research field is publication bias, which refers to the tendency that studies with significant studies more often are published, while those with insignificant results are disregarded (Joober et al., 2012; Rosenthal, 1979). Publication bias can thus create an erroneous impression of the evidence for a phenomenon. Besides the fact that null findings also have epistemic value, insignificant results encourage researchers to explore theoretical explanations that are not covered in existing theories. In turn, researchers may come up with novel hypotheses concerning the circumstances that influence the phenomena of interest.

Preregistration helps mitigate publication bias by rendering it more difficult to conceal that a study was planned. One strategy to assess the likelihood of publication bias within a line of research is funnel plot analyses conducted as part of meta-analyses. Funnel plot analyses show whether published studies with relatively high standard error tend to yield above-average results (Sterne et al., 2011). This is visualized as a negatively skewed distribution of results that thus conflicts with the assumption of normal distribution, for which publication bias is a possible explanation. When there are no funnel plot analyses nor tradition for preregistration, as there seemingly is not in the current field, the credibility of an evidence base should be called into question.

Recommendation 7: Be transparent about stakeholder interests and obligations

Our final recommendation is to be clearheaded about the larger context within which studies on CA-psychotherapy are conducted. This larger context includes a diverse set of stakeholders that share the overarching goal of developing an intervention for medical use. Among these stakeholders are, of course, scientists, clinicians, and practitioners, but also social institutions, including governments and public health administrators. And, finally, business interests, investors, and engineers. These stakeholders diverge in their legal and ethical obligations, as well as intentions. Social institutions aim to solve social and policy

issues, such as a looming mental health crisis. Their obligations are to the public, including their constituents, and their behavior is subject to laws that govern policy development and implementation. Scientists and medical professionals, on the other hand, have more narrow obligations, typically limited to their professions. Their behavior is governed by codes of ethics and laws about malpractice. Finally, business interests are often burdened with obligations to shareholders and investors and may see CA-psychotherapy primarily as a business opportunity. Their legal and ethical obligations will diverge from those that govern the behavior of scientists and medical professionals, as well as public institutions, even if their overarching goal is the same.

This diversity creates opportunities for science and medicine to have a significant impact on society. It also creates a danger of mixing obligations specific to one domain with those that burden other domains. For example, the short-term interests of investors for a successful CA-psychotherapy intervention could override the requirements of scientific rigor. Or public pressure on institutions to deliver solutions to social problems could re-direct funding to promising technological solutions at the expense of long-term investments in medical infrastructure and professionals. Arguably, what may help all stakeholders reach the overarching goal of delivering a successful CA-psychotherapy intervention given this context is to be transparent about their domain-specific obligations during the development and research into CA-psychotherapy interventions.

6. Conclusion

While research in CA-psychotherapy is flourishing these years, little emphasis has been placed on the therapeutic processes in these interventions. A deeper understanding of how CA-psychotherapy functions is crucial to improve these treatments and identify who will benefit from them, and who will not.

Our aim for this narrative review was to help pave the way for a process-oriented research direction in the realm of CA-psychotherapy. To achieve this, we conducted a narrative review of the literature and outlined the emerging research trends in this field. We identified three themes in the literature: relationship variables, emotional venting, and cognitive mechanisms. The overall conclusion drawn from this narrative review is that the epistemic value of the current evidence base is weak, and that several precautions should be made in future research. To help this, we presented a series of recommendations for future research, which we hope will help instigate methodologically sound investigations of various aspects of CA-psychotherapy.

However, besides the methodological challenges there are also fundamental problems associated with studying conversational agents. Particularly, there are basic assumptions of psychotherapy (e.g., that the client can form an affectionate bond with the therapist) that cannot be taken for granted in CA-psychotherapy. Consequently, future

research should not only aim to unravel the active ingredients of CA-psychotherapy by utilizing sound methods. Also, we should seek to identify and evaluate the extent to which basic assumptions of interpersonal interactions apply to human-computer interactions.

As a final comment, at this very moment we are potentially facing a paradigm shift in how mental health treatments are delivered due to the technological revolution currently taking place with conversational agents driven by large language models. While this transformation holds enormous potential for increasing accessibility to healthcare, it also implies that it has never been more important to study this subject to establish the potentials and boundaries of conversational agents.

7. References

A-Tjak, J. G. L., Morina, N., Topper, M., & Emmelkamp, P. M. G. (2021). One year follow-up and mediation in cognitive behavioral therapy and acceptance and commitment therapy for adult depression. *BMC Psychiatry*, *21*(1). https://doi.org/10.1186/s12888-020-03020-1

Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications*, *2*, 100006. https://doi.org/https://doi.org/10.1016/j.mlwa.2020.100006

Akiyoshi, T., Nakanishi, J., Ishiguro, H., Sumioka, H., & Shiomi, M. (2021). A Robot That Encourages Self-Disclosure to Reduce Anger Mood. *IEEE Robotics and Automation Letters*, *PP*, 1-1. https://doi.org/10.1109/LRA.2021.3102326

Al Farisi, R., Ferdiana, R., & Adji, T. B. (2022, 2022). The Effect of Anthropomorphic Design Cues on Increasing Chatbot Empathy.

Allouch, M., Azaria, A., & Azoulay, R. (2021). Conversational Agents: Goals, Technologies, Vision and Challenges. *Sensors (Basel)*, *21*(24). https://doi.org/10.3390/s21248448

Baier, A. L., Kline, A. C., & Feeny, N. C. (2020). Therapeutic alliance as a mediator of change: A systematic review and evaluation of research. *Clinical Psychology Review*, *82*, 101921. https://doi.org/https://doi.org/10.1016/j.cpr.2020.101921

Baldwin, S. A., & Imel, Z. E. (2013). Therapist effects: Findings and methods. *Bergin and Garfield's handbook of psychotherapy and behavior change*, *6*, 258-297.

Barrett-Lennard, G. (1993). The phases and focus of empathy. *The British journal of medical psychology*, *66 (Pt 1)*, 3-14. https://doi.org/10.1111/j.2044-8341.1993.tb01722.x

Beck, A. T. (1976). *Cognitive therapy and the emotional disorders*. International Universities Press.

Bell, E. C., Marcus, D. K., & Goodlad, J. K. (2013). Are the parts as good as the whole? A meta-analysis of component treatment studies. *Journal of consulting and clinical psychology*, *81*(4), 722-736. https://doi.org/10.1037/a0033004

Bohart, A. (1980). Toward a cognitive theory of catharsis. *Psychotherapy: Theory, Research & Practice*, *17*, 192-201. https://doi.org/10.1037/h0085911

Bohart, A. C., & Greenberg, L. S. (1997). Empathy and psychotherapy: An introductory overview. In *Empathy reconsidered: New directions in psychotherapy.* (pp. 3-31). American Psychological Association. https://doi.org/10.1037/10226-018

Bordin, E. S. (1979). The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, Research & Practice*, *16*(3), 252-260. https://doi.org/10.1037/h0085885

Bowlby, J. (1980). *Attachment and loss*. Basic Books.

Braun, S. R., Gregor, B., & Tran, U. S. (2013). Comparing Bona Fide Psychotherapies of Depression in Adults with Two Meta-Analytical Approaches. *PloS one*, *8*(6), e68135. https://doi.org/10.1371/journal.pone.0068135

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Daniel, Wu, J., Winter, C., . . . Amodei, D. (2020). Language Models are Few-Shot Learners. *arXiv pre-print server*. https://arxiv.org/abs/2005.14165

Chen, S.-C., Moyle, W., Jones, C., & Petsky, H. (2020). A social robot intervention on depression, loneliness, and quality of life for Taiwanese older adults in long-term care. *International Psychogeriatrics*, *32*(8), 981-991. https://doi.org/10.1017/S1041610220000459

Crits-Christoph, P., Baranackie, K., Kurcias, J. S., Beck, A. T., Carroll, K., Perry, K., Luborsky, L., McLellan, A., Woody, G., Thompson, L., Gallagher, D., & Zitrin, C. (1991). Meta-analysis of therapist effects in psychotherapy outcome studies. *Psychotherapy Research*, *1*(2), 81-91. https://doi.org/10.1080/10503309112331335511

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

Cuijpers, P., Cristea, I., Karyotaki, E., Reijnders, M., & Hollon, S. (2017). Component studies of psychological treatments of adult depression: A systematic review and meta-analysis. *Psychotherapy Research*, *29*, 1-15. https://doi.org/10.1080/10503307.2017.1395922

Cuijpers, P., Reijnders, M., & Huibers, M. J. H. (2019). The Role of Common Factors in Psychotherapy Outcomes. *Annu Rev Clin Psychol*, *15*, 207-231. https://doi.org/10.1146/annurev-clinpsy-050718-095424

David, D., Cristea, I., & Hofmann, S. G. (2018). Why Cognitive Behavioral Therapy Is the Current Gold Standard of Psychotherapy. *Frontiers in Psychiatry*, *9*, 4-4. https://doi.org/10.3389/fpsyt.2018.00004

de Gennaro, M., Krumhuber, E. G., & Lucas, G. (2020). Effectiveness of an empathic chatbot in combating adverse effects of social exclusion on mood. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.03061

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, *11*, 227-268. https://doi.org/10.1207/S15327965PLI1104_01

Elkins, D. N. (2009). The Medical Model in Psychotherapy:Its Limitations and Failures. *Journal of Humanistic Psychology*, *49*(1), 66-84. https://doi.org/10.1177/0022167807307901

Ellis-Brush, K. (2021). Augmenting Coaching Practice through digital methods. *International Journal of Evidence Based Coaching and Mentoring, Spec Iss 15*, 187-197. https://doi.org/10.24384/er2p-4857

Fenn, K., & Byrne, M. (2013). The key principles of cognitive behavioural therapy. *InnovAiT*, *6*(9), 579-585. https://doi.org/10.1177/1755738012471029

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

Frank, J. D., & Frank, J. B. (1991). *Persuasion and healing: A comparative study of psychotherapy, 3rd ed*. Johns Hopkins University Press.

Fulmer, R., Joerin, A., Gentile, B., Lakerink, L., & Rauws, M. (2018). Using Integrative Psychological Artificial Intelligence to Relieve Symptoms of Depression and Anxiety in Students (Preprint). *JMIR Mental Health*, *5*. https://doi.org/10.2196/mental.9782

Gambino, A., Fox, J., & Ratan, R. (2020). Building a Stronger CASA: Extending the Computers Are Social Actors Paradigm. *1*, 71-86. https://doi.org/10.30658/hmc.1.5

Garfield, S. L. (1995). *Psychotherapy: An eclectic-integrative approach, 2nd ed*. John Wiley & Sons.

Gelso, C. (2014). A tripartite model of the therapeutic relationship: Theory, research, and practice. *Psychotherapy Research*, *24*(2), 117-131. https://doi.org/10.1080/10503307.2013.845920

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

Hayes, S. C., Ciarrochi, J., Hofmann, S. G., Chin, F., & Sahdra, B. (2022). Evolving an idionomic approach to processes of change: Towards a unified personalized science of human improvement. *Behaviour Research and Therapy*, *156*, 104155. https://doi.org/https://doi.org/10.1016/j.brat.2022.104155

Hayes, S. C., Luoma, J. B., Bond, F. W., Masuda, A., & Lillis, J. (2006). Acceptance and commitment therapy: model, processes and outcomes. *Behav Res Ther*, *44*(1), 1-25. https://doi.org/10.1016/j.brat.2005.06.006

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*, e43862. https://doi.org/10.2196/43862

He, Y., Yang, L., Zhu, X., Wu, B., Zhang, S., Qian, C., & Tian, T. (2022). 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*, *24*(11), e40719. https://doi.org/10.2196/40719

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The Extent and Consequences of P-Hacking in Science. *PLOS Biology*, *13*(3), e1002106. https://doi.org/10.1371/journal.pbio.1002106

Horvath, A. O., Del Re, A. C., Flückiger, C., & Symonds, D. (2011). Alliance in individual psychotherapy. *Psychotherapy*, *48*(1), 9-16. https://doi.org/https://doi.org/10.1037/a0022186

Hurley, M. E., Lang, B. H., & Smith, J. N. (2023). Therapeutic Artificial Intelligence: Does Agential Status Matter? *The American Journal of Bioethics*, *23*(5), 33-35. https://doi.org/10.1080/15265161.2023.2191037

Ignatius, E., & Kokkonen, M. (2007). Factors contributing to verbal self-disclosure. *Nordic Psychology*, *59*(4), 362-391. https://doi.org/10.1027/1901-2276.59.4.362

Jeong, S., Alghowinem, S., Aymerich-Franch, L., Arias, K., Lapedriza, A., Picard, R., Hae, & Breazeal, C. (2020). A Robotic Positive Psychology Coach to Improve College Students' Wellbeing. *arXiv pre-print server*. https://doi.org/10.1109/RO-MAN47096.2020.9223588

Johannsen, M., Nissen, E. R., Lundorff, M., & O'Toole, M. S. (2022a). Mediators of acceptance and mindfulness-based therapies for anxiety and depression: A systematic review and meta-analysis. *Clin Psychol Rev*, *94*, 102156. https://doi.org/10.1016/j.cpr.2022.102156

Johannsen, M., Nissen, E. R., Lundorff, M., & O'Toole, M. S. (2022b). Mediators of acceptance and mindfulness-based therapies for anxiety and depression: A systematic review and meta-analysis. *Clinical Psychology Review*, *94*, 102156. https://doi.org/https://doi.org/10.1016/j.cpr.2022.102156

Johanson, D., Ahn, H., Lim, J., Lee, C., Sebaratnam, G., Macdonald, B., & Broadbent, E. (2020). Use of humor by a healthcare robot positively affects user perceptions and behavior. *Technology, Mind, and Behavior*, *1*. https://doi.org/10.1037/tmb0000021

Joober, R., Schmitz, N., Annable, L., & Boksa, P. (2012). Publication bias: What are the challenges and can they be overcome? *Journal of Psychiatry & Neuroscience*, *37*(3), 149-152. https://doi.org/10.1503/jpn.120065

Jørgensen, C. R. (2004). Active Ingredients in Individual Psychotherapy: Searching for Common Factors. *Psychoanalytic Psychology*, *21*(4), 516-540. https://doi.org/https://doi.org/10.1037/0736-9735.21.4.516

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/https://doi.org/10.1080/10447318.2022.2163775

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

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

Kazdin, A. E. (2007). Mediators and mechanisms of change in psychotherapy research. *Annu Rev Clin Psychol*, *3*, 1-27. https://doi.org/10.1146/annurev.clinpsy.3.022806.091432

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

Kirsch, I., Wampold, B., & Kelley, J. M. (2016). Controlling for the placebo effect in psychotherapy: Noble quest or tilting at windmills? *Psychology of Consciousness: Theory, Research, and Practice*, *3*, 121-131. https://doi.org/10.1037/cns0000065

Klos, M. C., Escoredo, M., Joerin, A., Lemos, V. N., Rauws, M., & Bunge, E. L. (2021). Artificial Intelligence–Based Chatbot for Anxiety and Depression in University

Students: Pilot Randomized Controlled Trial [Original Paper]. *JMIR Form Res*, *5*(8), e20678. https://doi.org/10.2196/20678

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174.

Lavelle, J., Dunne, N., Mulcahy, H. E., & McHugh, L. (2022). Chatbot-delivered cognitive defusion versus cognitive restructuring for negative self-referential thoughts: A pilot study. *The Psychological Record*, *72*(2), 247-261. https://doi.org/https://doi.org/10.1007/s40732-021-00478-7

Lee, J., Lee, D., & Lee, J.-G. (2022). Influence of Rapport and Social Presence with an AI Psychotherapy Chatbot on Users' Self-Disclosure. *International Journal of Human–Computer Interaction*, 1-12. https://doi.org/10.1080/10447318.2022.2146227

Lemmens, L. H. J. M., Galindo-Garre, F., Arntz, A., Peeters, F., Hollon, S. D., Derubeis, R. J., & Huibers, M. J. H. (2017). Exploring mechanisms of change in cognitive therapy and interpersonal psychotherapy for adult depression. *Behaviour Research and Therapy*, *94*, 81-92. https://doi.org/10.1016/j.brat.2017.05.005

Lemmens, L. H. J. M., Müller, V. N. L. S., Arntz, A., & Huibers, M. J. H. (2016). Mechanisms of change in psychotherapy for depression: An empirical update and evaluation of research aimed at identifying psychological mediators. *Clinical Psychology Review*, *50*, 95-107. https://doi.org/10.1016/j.cpr.2016.09.004

Lilienfeld, S., Lynn, S., & Ammirati, R. (2015). Science Versus Pseudoscience. In. https://doi.org/10.1002/9781118625392.wbecp572

Liu, H., Peng, H., Song, X., Xu, C., & Zhang, M. (2022). Using AI chatbots to provide self-help depression interventions for university students: A randomized trial of effectiveness. *Internet Interv*, *27*, 100495. https://doi.org/10.1016/j.invent.2022.100495

Liverant, G. I., Hofmann, S. G., & Litz, B. T. (2004). Coping and anxiety in college students after the September 11th terrorist attacks. *Anxiety, Stress & Coping: An International Journal*, *17*(2), 127-139. https://doi.org/10.1080/0003379042000221412

Longmore, R., & Worrell, M. (2007). Do we need to challenge thoughts in cognitive behavior therapy? *Clinical Psychology Review*, *27*, 173-187. https://doi.org/10.1016/j.cpr.2006.08.001

Lorenzo-Luaces, L. (2023). Identifying active ingredients in cognitive-behavioral therapies: What if we didn't? *Behaviour Research and Therapy*, *168*, 104365. https://doi.org/https://doi.org/10.1016/j.brat.2023.104365

Loveys, K., Hiko, C., Sagar, M., Zhang, X., & Broadbent, E. (2022). “I felt her company”: A qualitative study on factors affecting closeness and emotional support seeking with an embodied conversational agent. *International Journal of Human-Computer Studies, 160*, 102771. https://doi.org/https://doi.org/10.1016/j.ijhcs.2021.102771

Luborsky, L., Rosenthal, R., Diguer, L., Andrusyna, T. P., Berman, J. S., Levitt, J. T., Seligman, D. A., & Krause, E. D. (2002). The dodo bird verdict is alive and well--mostly. *Clinical Psychology: Science and Practice*, *9*, 2-12. https://doi.org/10.1093/clipsy.9.1.2

Lucas, G. M., Rizzo, A., Gratch, J., Scherer, S., Stratou, G., Boberg, J., & Morency, L.-P. (2017). Reporting Mental Health Symptoms: Breaking Down Barriers to Care with Virtual Human Interviewers [Original Research]. *Frontiers in Robotics and AI, 4*. https://doi.org/10.3389/frobt.2017.00051

Ly, K. H., Ly, A.-M., & Andersson, G. (2017). A fully automated conversational agent for promoting mental well-being: A pilot RCT using mixed methods. *Internet Interventions, 10*, 39-46. https://doi.org/https://doi.org/10.1016/j.invent.2017.10.002

McAleavey, A., & Castonguay, L. (2015). The Process of Change in Psychotherapy: Common and Unique Factors. In (pp. 293-310). https://doi.org/10.1007/978-3-7091-1382-0_15

Meng, J., & Dai, Y. (2021). Emotional Support from AI Chatbots: Should a Supportive Partner Self-Disclose or Not? *Journal of Computer-Mediated Communication*, *26*(4), 207-222. https://doi.org/10.1093/jcmc/zmab005

Mulder, R., Murray, G., & Rucklidge, J. (2017). Common versus specific factors in psychotherapy: opening the black box. *Lancet Psychiatry*, *4*(12), 953-962. https://doi.org/10.1016/S2215-0366(17)30100-1

Norcross, J. (2019). *Psychotherapy Relationships That Work: Evidence-Based Responsiveness*.

Norcross, J. C. (2002). *Psychotherapy Relationships That Work : Therapist Contributions and Responsiveness to Patients*. Oxford University Press. http://ebookcentral.proquest.com/lib/asb/detail.action?docID=281299

Pentina, I., Hancock, T., & Xie, T. (2023). Exploring relationship development with social chatbots: A mixed-method study of replika. *Computers in Human Behavior*, *140*, 107600. https://doi.org/https://doi.org/10.1016/j.chb.2022.107600

Powers, M. B., de Kleine, R. A., & Smits, J. A. J. (2017). Core Mechanisms of Cognitive Behavioral Therapy for Anxiety and Depression: A Review. *Psychiatric Clinics of*

North America, *40*(4), 611-623. https://doi.org/https://doi.org/10.1016/j.psc.2017.08.010

Pujiarti, R. N., Lee, B., & Yi, M. Y. (2022). Enhancing User's Self-Disclosure through Chatbot's Co-Activity and Conversation Atmosphere Visualization. *International Journal of Human–Computer Interaction*, *38*(18-20), 1891-1908. https://doi.org/10.1080/10447318.2022.2116414

Qian, Y., Tonya, N., Soravis, P., & Niloufar, S. (2019). *"I Almost Fell in Love with a Machine": Speaking with Computers Affects Self-disclosure* Glasgow, Scotland Uk. https://doi.org/10.1145/3290607.3312918

https://dl.acm.org/doi/pdf/10.1145/3290607.3312918

Quigley, L., Dozois, D. J. A., Bagby, R. M., Lobo, D. S. S., Ravindran, L., & Quilty, L. C. (2019). Cognitive change in cognitive-behavioural therapy v. pharmacotherapy for adult depression: a longitudinal mediation analysis. *Psychological Medicine*, *49*(15), 2626-2634. https://doi.org/10.1017/s0033291718003653

Rogers, C. R. (1951). *Client-centered therapy; its current practice, implications, and theory*. Houghton Mifflin.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results [doi:10.1037/0033-2909.86.3.638]. *86*, 638-641. https://doi.org/10.1037/0033-2909.86.3.638

Rosenzweig, S. (1936). SOME IMPLICIT COMMON FACTORS IN DIVERSE METHODS OF PSYCHOTHERAPY. *American Journal of Orthopsychiatry*, *6*(3), 412-415. https://doi.org/https://doi.org/10.1111/j.1939-0025.1936.tb05248.x

Røssberg, J. I., Evensen, J., Dammen, T., Wilberg, T., Klungsøyr, O., Jones, M., Bøen, E., Egeland, R., Breivik, R., Løvgren, A., & Ulberg, R. (2021). Mechanisms of change and heterogeneous treatment effects in psychodynamic and cognitive behavioural therapy for patients with depressive disorder: a randomized controlled trial. *BMC Psychology*, *9*(1). https://doi.org/10.1186/s40359-021-00517-6

Saxon, D., Firth, N., & Barkham, M. (2017). The Relationship Between Therapist Effects and Therapy Delivery Factors: Therapy Modality, Dosage, and Non-completion. *Administration and Policy in Mental Health and Mental Health Services Research*, *44*(5), 705-715. https://doi.org/10.1007/s10488-016-0750-5

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

Schuetzler, R. M., Giboney, J. S., Grimes, G. M., & Nunamaker, J. F., Jr. (2018). The influence of conversational agent embodiment and conversational relevance on socially desirable responding. *Decision Support Systems*, *114*, 94-102. https://doi.org/https://doi.org/10.1016/j.dss.2018.08.011

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

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*, 102601. https://doi.org/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*, 102903. https://doi.org/10.1016/j.ijhcs.2022.102903

Smedegaard, C. V. (2022). Novelty Knows No Boundaries: Why a Proper Investigation of Novelty Effects Within SHRI Should Begin by Addressing the Scientific Plurality of the Field [Perspective]. *Frontiers in Robotics and AI*, *9*. https://doi.org/10.3389/frobt.2022.741478

Soma, C. S., Baucom, B. R. W., Xiao, B., Butner, J. E., Hilpert, P., Narayanan, S., Atkins, D. C., & Imel, Z. E. (2020). Coregulation of therapist and client emotion during psychotherapy. *Psychotherapy Research*, *30*(5), 591-603. https://doi.org/10.1080/10503307.2019.1661541

Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., Carpenter, J., Rücker, G., Harbord, R. M., Schmid, C. H., Tetzlaff, J., Deeks, J. J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D. G., Moher, D., & Higgins, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*, *343*, d4002. https://doi.org/10.1136/bmj.d4002

ter Stal, S., Kramer, L. L., Tabak, M., op den Akker, H., & Hermens, H. (2020). Design Features of Embodied Conversational Agents in eHealth: a Literature Review. *International Journal of Human-Computer Studies*, *138*, 102409. https://doi.org/https://doi.org/10.1016/j.ijhcs.2020.102409

Trần, V., Szabó, Á., Ward, C., & Jose, P. E. (2023). To vent or not to vent? The impact of venting on psychological symptoms varies by levels of social support. *International Journal of Intercultural Relations*, *92*, 101750. https://doi.org/https://doi.org/10.1016/j.ijintrel.2022.101750

Wampold, B., Mondin, G., Moody, M., Stich, F., Benson, K., & Ahn, H.-n. (1997). A meta-analysis of outcome studies comparing bona fide psychotherapies: Empirically, "all must have prizes". *Psychological Bulletin - PSYCHOL BULL*, *122*, 203-215. https://doi.org/10.1037//0033-2909.122.3.203

Wampold, B. E. (2015). How important are the common factors in psychotherapy? An update. *World Psychiatry*, *14*(3), 270-277. https://doi.org/10.1002/wps.20238

Wampold, B. E. (2021). Healing in a Social Context: The Importance of Clinician and Patient Relationship. *Front Pain Res (Lausanne)*, *2*, 684768. https://doi.org/10.3389/fpain.2021.684768

Wampold, B. E., & Imel, Z. E. (2015). *The great psychotherapy debate: The evidence for what makes psychotherapy work, 2nd ed*Routledge/Taylor & Francis Group.

Wang, J., Lloyd-Evans, B., Marston, L., Mann, F., Ma, R., & Johnson, S. (2020). Loneliness as a predictor of outcomes in mental disorders among people who have experienced a mental health crisis: a 4-month prospective study. *BMC Psychiatry*, *20*(1). https://doi.org/10.1186/s12888-020-02665-2

Watson, J. C. (2016). THE ROLE OF EMPATHY IN PSYCHOTHERAPY

THEORY, RESEARCH, AND PRACTICE. In D. J. Cain, K. Keenan, & S. Rubin (Eds.), *Humanistic Psychotherapies* (pp. 115-146). American Psychological Association. http://www.jstor.org.ez.statsbiblioteket.dk/stable/j.ctv1chrwdk.10

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, *9*(1), 36–45. https://doi.org/10.1145/365153.365168

Wieland, L. C. (2023). Relational Reciprocity from Conversational Artificial Intelligence in Psychotherapy. *The American Journal of Bioethics*, *23*(5), 35-37. https://doi.org/10.1080/15265161.2023.2191033

Wills, T. A., & Shinar, O. (2000). Measuring perceived and received social support. In *Social support measurement and intervention: A guide for health and social scientists.* (pp. 86-135). Oxford University Press. https://doi.org/10.1093/med:psych/9780195126709.003.0004

Wilson, A. L. (1993). The promise of situated cognition. *New Directions for Adult and Continuing Education*, *1993*(57), 71-79. https://doi.org/10.1002/ace.36719935709

Yalom, I. D. (1989). *Love's executioner: And other tales of psychotherapy*. Basic Books.

Yi-Chieh, L., Naomi, Y., & Yun, H. (2020). Designing a Chatbot as a Mediator for Promoting Deep Self-Disclosure to a Real Mental Health Professional. *Proc. ACM Hum.-Comput. Interact.*, *4*(CSCW1), Article 31. https://doi.org/10.1145/3392836

Yi-Chieh, L., Naomi, Y., Yun, H., & Wai, F. (2020). "I Hear You, I Feel You": Encouraging Deep Self-disclosure through a Chatbot. 1–12. https://doi.org/10.1145/3313831.3376175

Figure 1.

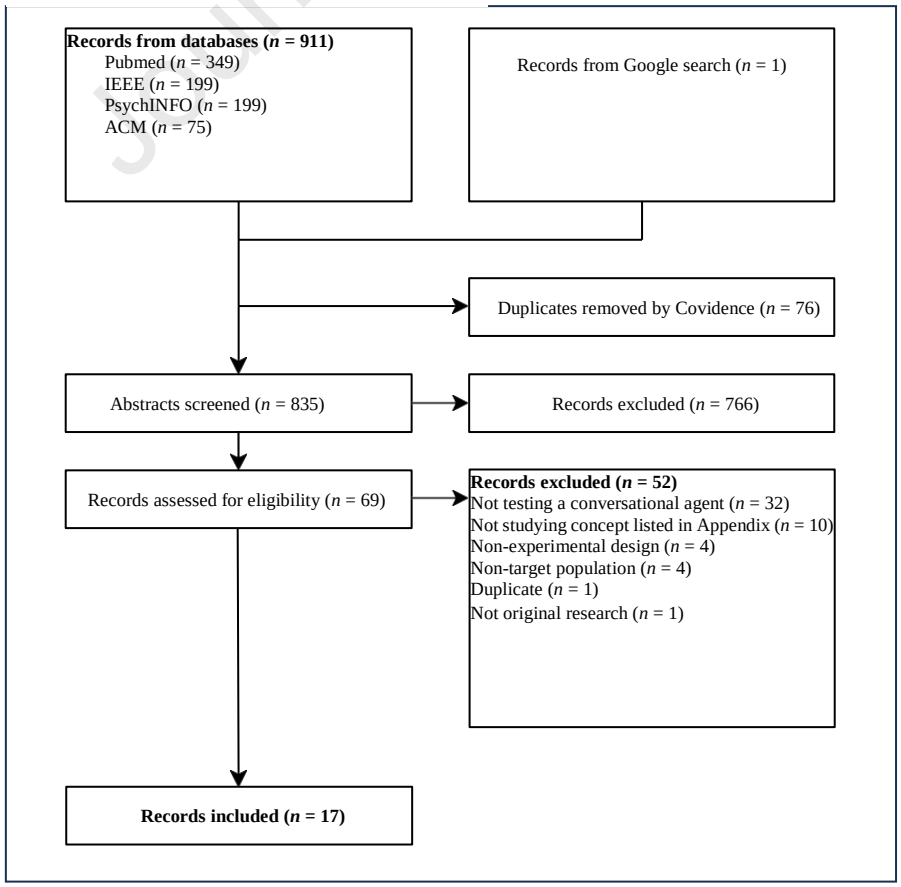

Table 1: Overview of included records

Author(s)	Sample	Condition(s)	Intervention	Main measurement(s)	Main outcome(s)
Relationship variables: 7 studies					
Johanson et al. (2020)	91 university students and employees (M_{age} = 25.03) (SD_{age} = 11.06) ($Range_{age}$ = 17-63).	Randomized between-condition experiment: • Humorous robot. • Neutral robot. The social robot communicated by listening, speaking, and facial expressions.	Single-session conversation on flu vaccination. A priori testing was conducted to ensure that the robot was perceived as humorous.	McGill Friendship Questionnaire. Consultation and Relational Empathy Questionnaire.	A significant condition-time effect was observed in empathy (p = .02, n_p^2 = .06). Main effect of condition at post-intervention is not reported in the paper.
Jeong et al. (2020)	35 students.	Pretest-posttest experimental design: • Social robot with a robotic appearance.	Seven in-home sessions each with a duration of 3-6 minutes in which the social robot delivered a positive psychology intervention, such as expressing gratitude, expressive writing, and identifying strengths.	Working Alliance Inventory-Short Revised.	The working alliance was rated as M = 3.43, SD = 0.83 at post-intervention. *Note:* Inferential analyses of the relationship to therapy outcomes were not reported.
de Gennaro et al. (2020)	128 students (M_{age} = 24.12) (SD_{age} = 5.91)	Between-condition experiment: • Empathetic chatbot • Interactive questionnaire.	The experiment encompassed two phases. First, A social exclusion	Mood as assessed with The Positive and Negative Affect Schedule.	Social exclusion decreased positive affect (p < .001, d = 0.8) and increased negative affect (p < .001, d = 0.42). Overall mood was rated as

Author(s)	Sample	Condition(s)	Intervention	Main measurement(s)	Main outcome(s)
		Note: besides the rule-based text communication, most of the communication relied on predefined multiple-choice answers.	experience, where the participants experienced receiving the fewest 'likes' for a personal description on a fictive social media platform. Second, self-disclosure to an empathetic chatbot (e.g., "I am sorry this happened to you") or an interactive questionnaire (e.g., "Thank you for letting us know").		significantly better after disclosing to the empathetic chatbot compared to the interactive questionnaire (p < .001, d = 0.37) with feeling of social exclusion as a covariate.
Ellis-Brush (2021)	48 volunteers.	Pretest-posttest experiment: Wysa therapy chatbot.	CBT across 8 weeks.	The Working Alliance Inventory.	No significant within-condition increase in working alliance between 1 week and 9 weeks.
Al Farisi et al. (2022)	30 students.	Between-condition experiment: • Empathetic chatbot. • Informational chatbot. Note: Unclear whether randomized allocation was performed.	A single service conversation with a chatbot that functioned as a support service for university students. The empathetic chatbot used the 'I' pronoun, expressed emotions, engaged	Ad hoc self-report scale to assess empathy.	The empathetic chatbot was rated as significantly more empathetic than the informative chatbot (p < .001).

Author(s)	Sample	Condition(s)	Intervention	Main measurement(s)	Main outcome(s)
			in small-talking, and asked follow-up questions.		
Liu et al. (2022)	63 students with moderate symptoms of depression were analyzed. 83 students were enrolled (M_{age} = 23.08) (SD_{age} = 1.76) ($Range_{age}$ = 19-28).	Randomized controlled trial: • XiaoNan chatbot. • Bibliotherapy.	16 weeks chatbot-delivered CBT vs. e-book on depression.	The Working Alliance Inventory-Short Revised.	The working alliance with the chatbot was rated as significantly better than with the e-book (p < .001, d = 1.85).
He et al. (2022)	148 students (after 12.8% drop-out) with an average score in depression (M_{age} = 18.78) (SD_{age} = 0.89) ($Range_{age}$ = 17-21).	Randomized controlled trial: • XiaoE therapy chatbot. • Xoaoai general-purpose chatbot. • Bibliotherapy.	1-week study period. The therapy chatbot delivered CBT.	Working Alliance Questionnaire.	A main effect of condition on working alliance was observed (p = .04) with higher ratings in the therapy chatbot condition (M = 53.94, SD = 5.96) than in the general-purpose chatbot condition (M = 50.68, SD = 6.87) and e-book condition (M = 50.35, SD = 9.38).
Emotional venting: 9 studies					

Author(s)	Sample	Condition(s)	Intervention	Main measurement(s)	Main outcome(s)
Schuetzler et al. (2018)	165 students after exclusion of 33 students (M_{age} = 20.8) (SD_{age} = 1.6).	Randomized between-condition experiment: face-to-face vs. online survey vs chatbot (embodiment: yes vs. no; relevant follow-questions: yes vs. no).	In a single session, the conversational agents asked questions about the participants' daily life, including their study and recreational activities.	Drinking behaviors (i.e., drinks per week and days since toxification). Health behaviors (i.e., daily vegetables and exercise.	Participants were significantly more likely to report drinking behaviors to CAs than to the human interviewer (p < .05, b = 0.92). Participants were significantly more likely to report drinking behaviors to CAs that asked relevant follow-up questions (p < .05, b = 0.79). Males reported significantly more drinking behaviors to CAs than women did (p < .001, b = 0.67).
Qian et al. (2019)	Convenience sample (N = 30) ($Range_{age}$ = 18-34).	Randomized between-condition experiment: 2 (reading vs. typing) * 2 (male vs. female voice) * 2 (listening vs. reading).	In a single session, the conversational agents asked increasingly more personal questions to the participants.	Word count. The number of questions answered. Sentiment analysis.	Assigned condition did significantly predict self-disclosures.
Yi-Chieh, Naomi and Yun (2020)	Students (N = 47).	Randomized between-condition experiment: Non vs. Low vs. High reciprocal self-disclosing chatbot.	Across three weeks, participants talked daily with a chatbot for 10 minutes about themselves, including answering sensitive questions, journaling, and small talking. After two weeks,	Qualitative coding transcriptions in terms of information, thoughts, and emotions using a standardized tool.	No significant differences in self-disclosure to the chatbot after the participants were informed that a health professional would receive their disclosed information. Self-disclosing chatbots elicited significantly more self-disclosures of feelings from sensitive questions than non-self-disclosing chatbots.

Author(s)	Sample	Condition(s)	Intervention	Main measurement(s)	Main outcome(s)
			they were informed that they could choose to share past and upcoming transcriptions of their disclosures with a mental health professional.		Highly self-disclosing chatbots elicited significantly more self-disclosures of feelings from journaling questions than less self-disclosing chatbots.
Yi-Chieh, Naomi, et al. (2020)	47 students.	Randomized between-condition experiment: Non vs. Low vs. High reciprocal self-disclosing chatbot.	Across three weeks, participants talked daily with a chatbot for 10 minutes about themselves, including answering sensitive questions, journaling, and small talking.	Self-reported trust, intimacy, and enjoyment.	The high-self-disclosing chatbot was rated as significantly more enjoyable than the non-self-disclosing chatbot ($p < .05$). The low-self-disclosing chatbot was rated as significantly less trustworthy than each of the other two chatbots ($p < .05$). After three weeks, the high-self-disclosing chatbot was rated significantly higher in intimacy than non-self-disclosing chatbots ($p < .01$).
Akiyoshi et al. (2021)	31 participants enrolled through employment agency (26 responses analyzed) (M_{age} = 41.1) (SD_{age} = 9.78)	Gender stratified between-condition experiment: • Zoomorphic robot ('Sota') using the column method from CBT to encourage self-	In single 7-10 minutes conversations, the robot 'Sota' used different strategies to elicit self-disclosures on thoughts, emotions, and self-schemes.	Qualitative coding and counting of self-disclosures (e.g., personal information such as hobbies and experiences) as opposed to non-self-disclosures	Significantly more self-disclosures to the robot combining the column method and extra conversational strategies compared to the robot only using the column method ($p < .001$, $r = .73$). Significant reduction in anger

Author(s)	Sample	Condition(s)	Intervention	Main measurement(s)	Main outcome(s)
		disclosures. • 'Sota' combines the column method with other conversational strategies to promote more self-disclosures.	In the combined condition, Sota was asking into the triggers of thoughts, asking for confirmation that the robot understood the participant correctly, and exploring the participants' self-schemes.	(e.g., mundane topics such as the weather). The Japanese version of the Profile of Mood States for measuring emotional states.	(p = .028, r = .53) from before to after interacting with the robot. No significant within-condition differences in other emotions, such as anxiety and depression.
Meng and Dai (2021)	211 students (M_{age} = 20.4) (SD_{age} = 2.28).	Randomized between-condition experiment: 2 (chatbot vs. human) * emotional support (yes vs. no) * reciprocal self-disclosure (yes vs. no).	In a single online session, participants chatted with the conversation partner (human or chatbot) about a stressful. Issue. The conversation partner asked questions to elicit self-disclosures in the participants.	Perceived Stress Scale. Single item assessing worry. Adapted items to assess perceived supportiveness of conversation partner.	Moderated mediation analyses showed that emotional support more strongly predicted stress reduction (*index* = .08, *SE* = 0.05, 95% CI [0.003, 0.19] and worry (*index* = .17, *SE* = 0.10, 95% CI [0.02, 0.39] through perceived emotional support in the human conditions than in the chatbot conditions. Notably, a three-way ANCOVA tests with neuroticism as control variable and worry as dependent variable showed no significant interaction, revealing that there was no combined effects of emotional support and reciprocal self-disclosure between the chatbot and human conditions (p = .23). See source for more

Author(s)	Sample	Condition(s)	Intervention	Main measurement(s)	Main outcome(s)
					outcomes.
Pujiarti et al. (2022)	87 students (M_{age} = 23.88)	Randomized between-condition experiment: • Chatbot. • Chatbot with conversation atmosphere visualization. • Chatbot with co-activity. • Chatbot (CAV and co-activity).	Participants chatted with the chatbot for 10 to 15 minutes each day for 10 days. The conversations included small talk and personal questions about race, religion, cheating, lies, personality, and body perception.	Word count. Sentiment analysis. Qualitative coding of information, thoughts, and emotions, as assessed using a standardized tool.	Conversation atmosphere visualization increased the disclosure of information on day 5 (p = .017, n^2 = .066) and 10 10 (p = .022, n^2 = .062) compared to no chatbot only. Co-activity increased an aggregate variable of information, thoughts, and feelings at day 5 (p = .02, n^2 = .063).
Lee et al. (2022)	303 nonclinical participants (33% in their 30s) (28.1% in their 40s) (19.5% in their 20s).	Experimental study: • Therapy chatbot.	At least 5-minutes of conversation with a therapy chatbot about current personal issues.	Self-report of self-disclosure.	The perceived social presence of chatbots predicted less self-disclosure (β = 0.80, p < .001). Fear of negative evaluation from chatbot predicted less self-disclosure (β = 0.40, p < .001). The interaction between social presence and fear of negative evaluation was a negative significant predictor of self-disclosure (β = -0.015, p < .001).
Kang and Kang (2023)	135 Students (M_{age} = 23.9) (SD_{age} = 3.02).	Randomized between-condition experiment: 2 (introvert vs.	10 to 20 minutes interview about mood, distress,	Self-report of honesty and depth of self-disclosures.	Honesty significantly decreased with the chatbot with the visual interface (p =

Author(s)	Sample	Condition(s)	Intervention	Main measurement(s)	Main outcome(s)
		extrovert chatbot) * 2 (male vs. female chatbot) * 2 (embodied vs. disembodied chatbot) * 2 (male vs. female participant).	relationships, and health.		.036) A significant interaction between the gender of the chatbot, the gender of the user, and the anthropomorphic design (p = .016). When there is a visual interface, honesty is highest toward the opposite gender agents. Females' depth of disclosures was significantly larger toward male chatbots (p = .003). See source for more outcomes.
Cognitive factors: 1 study					
Lavelle et al. (2022)	223 students enrolled. 161 participants were lost to follow-up or excluded. 28 participants were included for analyses. (M_{age} = 28.01) (SD_{age} = 10.29) ($Range_{age}$ = 18-68)* *Data for the 223 students enrolled.	Randomized controlled trial: • Chatbot delivering cognitive restructuring techniques. • Chatbot delivering cognitive defusion techniques. • Measurement only.	The two interventions consisted of five sessions each with a duration of about 10 minutes.	Self-report of believability, discomfort, negativity, and extremity of a self-chosen negative self-referential thought. Acceptance and Action Questionnaire-II. Cognitive Fusion Questionnaire-7. Positive and Negative Affect Schedule.	No significant between or within-condition effects. Note: a significant post-intervention difference in positive affect was found between the cognitive restructuring condition and measurement only condition, however, this difference also existed prior to the intervention.

Table 2: Main findings summarized as bullet points

Theme	Main findings	Sources
Relationship variables	Current research neither supports nor contests the possibility that the alliance serves a function in CA-psychotherapy.	Jeong et al. (2020) Ellish-Brush et al. (2021) Al Farisi et al. (2022) He et al. (2022) de Gennaro et al. (2020)
Emotional venting	Mixed evidence for the emotional outcomes of self-disclosing to conversational agents.	de Gennaro et al. (2020) Akiyoshi et al. (2021) Meng and Dai (2021)
	The tendency to self-disclose to conversational agents is moderated by several variables, such as gender, conversational strategy, and perceived social presence and fear of negative social evaluation of the agent.	Qian et al. (2019) Yi-Chieh et al. (2020) Schuetzler et al. (2018) Akiyoshi et al. (2021) Pujiarti et al. (2022) Lee et al. (2022)
Cognitive techniques	Current research neither supports nor contests the possibility that cognitive restructuring and defusion techniques are effective in CA-psychotherapy.	Lavelle et al. (2022)

Table 3: Future directions

	Recommendation	**Implementation**
Internal validity	Minimize threats to internal validity in process-research.	Employ time-lagged mediation analyses and component study designs.
	Utilize the accessibility of conversational agents to increase statistical power.	Aim toward sample sizes based on a priori power analyses.
External validity	Be aware of the multifaceted nature of conversational agents.	Be cautious about generalizing research findings based on one sort of conversational agent to another.
	Take the diversity of human reference groups into account.	Consider inter-individual differences and their implications for the generalizability of research utilizing human comparison conditions.
	Assess clinical populations.	Examine clinical populations for whom the intervention was designed to alleviate a hypothesized psychological deficit.
Research credibility	Preregister your studies to increase the credibility of research findings.	Refer to online preregistration tools such as the OSF registry and Aspredicted.
	Be transparent about stakeholder interests and obligations.	When interpreting research findings on CA-psychotherapy, be aware of potential industrial collaboration partners in the study and their interest.

Appendix

List of search terms derived from theoretical and review papers

1. Alliance (working alliance, therapeutic alliance, therapeutic relationship, therapeutic collaboration)
2. Empathy (reflective functioning, mentalization)
3. Exposure (desensitization, behavioral activation)
4. Cognitive decentering (defusion, distancing, disidentification, detachment, mindfulness, insight, self-understanding, metacognition, self-as-context, non-reactivity)
5. Self-compassion (experiential avoidance, psychological flexibility)
6. Cognitive restructuring (reappraisal, perspective change, dysfunctional assumptions, dysfunctional thinking, automatic thoughts, catastrophic thinking)
7. Self-disclosure (emotional venting)
8. Corrective emotional experience
9. Expectations (placebo)
10. Emotion regulation (coping)

Search strings

PsychInfo (abstract, title, keywords)

Ab,ti,su(("conversational agent*" OR "chatbot*" OR "dialogue system*" OR "virtual agent*" OR "virtual avatar*" OR "digital human*" OR "relational agent*" OR "computer agent*" OR "social robot*") **AND** (“active ingredient*” OR "mechanism of change" OR "mechanisms of change" OR "mechanism of action" OR "mechanisms of action" OR "change mechanism*" OR "working alliance" OR "therapeutic alliance" OR "therapeutic relationship" OR "therapeutic collaboration" OR "bond" OR "empath*" OR "reflective function*" OR "mentalization" OR "exposure" OR "desensitization" OR "behavioral activation" OR "decentering" OR "defusion" OR "distancing" OR "disidentification" OR "detachment" OR "mindfulness" OR "insight" OR "self-understanding" OR "meta-cognition" OR "self-as-context" OR "non-reactivity" OR "self-compassion" OR "experiential avoidance" OR "psychological flexibility" OR "cognitive restructuring" OR "reappraisal" OR "perspective change" OR "dysfunctional assumption*" OR "dysfunctional thinking" OR "automatic thought*" OR "self-disclosure" OR “emotional venting” OR "corrective emotional experience*" OR "emotion regulation" OR "coping" OR "catastrophic thinking" OR "expecta*" OR "placebo"))

Pubmed (abstract, title, MeSH terms)

("conversational agent*" OR "chatbot*" OR "dialogue system*" OR "virtual agent*" OR "virtual avatar*" OR "digital human*" OR "relational agent*" OR "computer agent*" OR

"social robot*") **AND** (“active ingredient*” OR "mechanism of change" OR "mechanisms of change" OR "mechanism of action" OR "mechanisms of action" OR "change mechanism*" OR "working alliance" OR "therapeutic alliance" OR "therapeutic relationship" OR "therapeutic collaboration" OR "bond" OR "empath*" OR "reflective function*" OR "mentalization" OR "exposure" OR "desensitization" OR "behavioral activation" OR "decentering" OR "defusion" OR "distancing" OR "disidentification" OR "detachment" OR "mindfulness" OR "insight" OR "self-understanding" OR "meta-cognition" OR "self-as-context" OR "non-reactivity" OR "self-compassion" OR "experiential avoidance" OR "psychological flexibility" OR "cognitive restructuring" OR "reappraisal" OR "perspective change" OR "dysfunctional assumption*" OR "dysfunctional thinking" OR "automatic thought*" OR "self-disclosure" OR “emotional venting” OR "corrective emotional experience*" OR "emotion regulation" OR "coping" OR "catastrophic thinking" OR "expecta*" OR "placebo")

ACM (abstract, title, author keywords)

("conversational agent*" OR "chatbot*" OR "dialogue system*" OR "virtual agent*" OR "virtual avatar*" OR "digital human*" OR "relational agent*" OR "computer agent*" OR "social robot*") **AND** (“active ingredient*” OR "mechanism of change" OR "mechanisms of change" OR "mechanism of action" OR "mechanisms of action" OR "change mechanism*" OR "working alliance" OR "therapeutic alliance" OR "therapeutic relationship" OR "therapeutic collaboration" OR "bond" OR "empath*" OR "reflective function*" OR "mentalization" OR "exposure" OR "desensitization" OR "behavioral activation" OR "decentering" OR "defusion" OR "distancing" OR "disidentification" OR "detachment" OR "mindfulness" OR "insight" OR "self-understanding" OR "meta-cognition" OR "self-as-context" OR "non-reactivity" OR "self-compassion" OR "experiential avoidance" OR "psychological flexibility" OR "cognitive restructuring" OR "reappraisal" OR "perspective change" OR "dysfunctional assumption*" OR "dysfunctional thinking" OR "automatic thought*" OR "self-disclosure" OR “emotional venting” OR "corrective emotional experience*" OR "emotion regulation" OR "coping" OR "catastrophic thinking" OR "expecta*" OR "placebo")

IEEE (all metadata)

(("All Metadata":"conversational agent") OR ("All Metadata":"conversational agents") OR ("All Metadata":"chatbot") OR ("All Metadata":"chatbots") OR ("All Metadata":"dialogue system") OR ("All Metadata":"dialogue systems") OR ("All Metadata":"virtual agent") OR ("All Metadata":"virtual agents") OR ("All Metadata":"virtual avatar") OR ("All Metadata":"virtual avatars") OR ("All Metadata":"digital human") OR ("All Metadata":"digital humans") OR ("All Metadata":"relational agent") OR ("All Metadata":"relational agents") OR ("All Metadata":"social robot") OR ("All Metadata":"social

robots") OR ("All Metadata":"social robotic") OR ("All Metadata":"social robotics")) **AND** (("All Metadata":"active ingredient") OR("All Metadata":"mechanism of change") OR ("All Metadata":"mechanisms of change") OR ("All Metadata":"mechanism of action") OR ("All Metadata":"mechanisms of action") OR ("All Metadata":"change mechanism") OR ("All Metadata":"change mechanisms") OR ("All Metadata":"working alliance") OR ("All Metadata":"therapeutic alliance") OR ("All Metadata":"therapeutic relationship") OR ("All Metadata":"therapeutic collaboration") OR ("All Metadata":"bond") OR ("All Metadata":"empathethic") OR ("All Metadata":"empathy") OR ("All Metadata":"reflective function") OR ("All Metadata":"reflective functioning") OR ("All Metadata":"mentalization") OR ("All Metadata":"exposure") OR ("All Metadata":"desensitization") OR ("All Metadata":"behavioral activation") OR ("All Metadata":"decentering") OR ("All Metadata":"defusion") OR ("All Metadata":"distancing") OR ("All Metadata":"disidentification") OR ("All Metadata":"detachment") OR ("All Metadata":"mindfulness") OR ("All Metadata":"insight") OR ("All Metadata":"self-understanding") OR ("All Metadata":"meta-cognition") OR ("All Metadata":"self-as-context") OR ("All Metadata":"non-reactivity") OR ("All Metadata":"self-compassion") OR ("All Metadata":"experiential avoidance") OR ("All Metadata":"psychological flexibility") OR ("All Metadata":"cognitive restructuring") OR ("All Metadata":"reappraisal") OR ("All Metadata":"perspective change") OR ("All Metadata":"dysfunctional assumption") OR ("All Metadata":"dysfunctional assumptions") OR ("All Metadata":"dysfunctional thinking") OR ("All Metadata":"automatic thought") OR ("All Metadata":"automatic thoughts") OR ("All Metadata":"self-disclosure") OR ("All Metadata":"emotional venting") OR ("All Metadata":"corrective emotional experience") OR ("All Metadata":"corrective emotional experiences") OR ("All Metadata":"emotion regulation") OR ("All Metadata":"coping") OR ("All Metadata":"catastrophic thinking") OR ("All Metadata":"expecta*") OR ("All Metadata":"placebo”

Declaration of interests

☒ 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.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: