Therapeutic Chatbots as Cognitive-Affective Artifacts

J. P. Grodniewicz[1] · **Mateusz Hohol**[1]

Abstract

Conversational Artificial Intelligence (CAI) systems (also known as AI "chatbots") are among the most promising examples of the use of technology in mental health care. With already millions of users worldwide, CAI is likely to change the landscape of psychological help. Most researchers agree that existing CAIs are not "digital therapists" and using them is not a substitute for psychotherapy delivered by a human. But if they are not therapists, what are they, and what role can they play in mental health care? To answer these questions, we appeal to two well-established and widely discussed concepts: cognitive and affective artifacts. Cognitive artifacts are artificial devices contributing functionally to the performance of a cognitive task. Affective artifacts are objects which have the capacity to alter subjects' affective state. We argue that therapeutic CAIs are a kind of cognitive-affective artifacts which contribute to positive therapeutic change by (i) simulating a (quasi-)therapeutic interaction, (ii) supporting the performance of cognitive tasks, and (iii) altering the affective condition of their users. This sheds new light on why virtually all existing mental health CAIs implement principles and techniques of Cognitive Behavioral Therapy — a therapeutic orientation according to which affective change and, ultimately, positive therapeutic change is mediated by cognitive change. Simultaneously, it allows us to conceptualize better the potential and limitations of applying these technologies in therapy.

Keywords Cognitive Artifacts · Affective Artifacts · Conversational Artificial Intelligence · Mental Health Chatbots · Psychotherapy · Simulation

1 Introduction

Conversational Artificial Intelligence systems (also known as "chatbots") are dialogue systems designed to engage in real-time conversational exchanges with human users. CAIs are already quite commonly used in fields like banking or e-commerce. More recently, they started becoming increasingly popular in mental health care (Abd-Alrazaq et al. 2020; Ahmed et al. 2023; Boucher et al. 2021; He et al. 2023; Laranjo et al. 2018; Vaidyam et al. 2019, Grodniewicz and Hohol 2023a).

Mental health chatbots are designed for the "delivery of therapy, training, and screening" (Abd-Alrazaq et al. 2019, p. 6). Many authors have already argued convincingly that — despite how they are often advertised by their producers and even described in some peer-reviewed papers (e.g., Fitzpatrick et al. 2017) — existing CAIs should not be considered therapists and the services they provide fall short of fully fledged psychotherapy. Firstly, it is unclear whether we can speak about the therapeutic relationship between chatbots and their clients/patients (Brown and Halpern 2021; Tekin 2023; Tong et al. 2022; Vagwala and Asher 2023, Grodniewicz and Hohol 2023a), and if we cannot, whether there can be psychotherapy without a therapeutic relationship (Safran and Muran 2006; Zilcha-Mano 2017). Secondly, they cannot engage in normal human discursive practices: they do not understand the concepts they use, and they cannot give reasons for their own actions (Sedlakova and Trachsel 2023). Thirdly, CAIs are currently able to deal with only narrowly construed, well-delineated tasks. They do not possess the full range of therapeutic skills, nor are they able to engage in conversation about all topics relevant to

✉ J. P. Grodniewicz
j.grodniewicz@gmail.com

✉ Mateusz Hohol
mateusz.hohol@uj.edu.pl

1 Copernicus Center for Interdisciplinary Studies, Jagiellonian University in Krakow, Krakow, Poland

the mental suffering of their clients (Laranjo et al. 2018; Pandey et al. 2022, Grodniewicz and Hohol 2023a).

However, if mental health chatbots are not therapists, what are they? Sedlakova and Traschel (2023) suggest the following preliminary answer to this question:

> CAI should not be understood as a tool merely implementing evidence-based therapies nor as a digital therapist, but as *a new artifact* that can change our interactions, concepts, epistemic field, and normative requirements and whose status on the spectrum between a tool and a therapist or an agent respectively, needs to be defined. (p. 3, *emphasis ours*)

In this paper, we undertake the task of defining the status of this new artifact.[1] We do it by appealing to two well-established and widely discussed concepts of *cognitive artifact* introduced by Norman (1991) and further investigated by such authors as Hutchins (1999); Brey (2005); Heersmink (2013); Fasoli (2018a), and *affective artifact* introduced by Piredda (2020), and further developed by, e.g., Heersmink (2021) and Viola (2021). Roughly, cognitive artifacts are artificial devices contributing to the performance of cognitive tasks, while affective artifacts are objects with the capacity to alter subjects' affective states (see further discussion in Sect. 3). Crucially, as has already been noted by Heersmink (2021), the same object can be a cognitive and affective artifact at the same time. Thus, we argue that mental health chatbots are a subtype of cognitive-affective artifact, which help their users achieve positive therapeutic change by (i) simulating a (quasi-)therapeutic interaction, (ii) supporting the performance of cognitive tasks, and (iii) altering the affective condition of the users.

The paper proceeds as follows. In the next section (Sect. 2), we briefly introduce the field of mental health chatbots focusing on the fact that the most popular among them implement the principles of so-called Cognitive Behavioral Therapy (CBT) — psychotherapeutic orientation, according to which the affective change, and, ultimately, the positive therapeutic change is mediated by cognitive change. In Sect. 3, we introduce and discuss the concepts of cognitive and affective artifacts. In Sect. 4, we demonstrate that chatbots can be both cognitive and affective artifacts. Finally, in Sect. 5, we argue

that therapeutic chatbots are a special kind of cognitive-affective artifact.

2 Mental Health Chatbots and Cognitive Behavioral Therapy

There are already dozens of available mental health chatbots used by millions of users worldwide. Notably, all the most popular products of this type, such as Woebot, Wysa, Youper, or Tess, are advertised as implementing the principles and techniques of CBT. CBT is one of the most popular modern therapeutic modalities (Cook et al. 2010; Prochaska and Norcross 2018) considered by many the gold standard of talk therapy (David et al. 2018). It is rooted in two venerable traditions in clinical psychology: *behavioral*, drawing from the work of such figures as Joseph Wolpe, Ivan Pavlov, Clark Hull, and B. F. Skinner, and *cognitive*, launched by Aaron Beck and Albert Ellis (O'Donohue and Ferguson 2016). Even though modern CBT incorporates behavioral processes, such as exposure, it is fair to say that its dominant narrative about the nature of psychopathology is cognitive. According to *the cognitive model of psychopathology*, "self-relevant thoughts, evaluations, and beliefs are key contributors to the development and persistence of psychopathological states" (Clark 2013, p. 1). To put it simply, mental suffering is caused by what we think about the world, ourselves, and others (the three are sometimes referred to as the cognitive triad (Beck 1967)).

More specifically, in a recent review of advances in cognitive theory and therapy, we read:

> What transforms normal adaptive reactions to disorders? We believe that the cause is faulty information processing… When information processing provides faulty information, other systems (e.g., affective, motivational, behavioral) no longer function in an adaptive way. Errors can result in other cognitive biases (e.g., interpretation, attention, memory), excessive or inappropriate affect, and maladaptive behavior. (Beck and Haigh 2014, p. 4)

Maladaptive patterns of information processing cause further affective, behavioral, and motivational problems, and ultimately, may give rise to mental disorders. This assumption has important consequences for psychotherapeutic treatment. According to CBT, therapeutic intervention aims at modifying maladaptive thinking patterns, i.e., cognitive change. It is the cognitive change, or so it is assumed, that mediates the change in patients' affect, and leads to positive therapeutic change measured by symptom reduction or

---

[1] Even though Sedlakova and Trachsel (2023) distinguish between tools and artifacts, they do not offer clear-cut definitions of either, so it is not obvious whether the two categories have to be mutually exclusive. We are not going to settle this issue in the present paper. Rather, we will specify the artifactual nature of therapeutic chatbots, leaving aside whether it simultaneously disqualifies them as "tools."

improvement in patients' overall well-being (Dobson and Dozois 2010; Lorenzo-Luaces et al. 2015).

But by what means is the cognitive change achieved during psychotherapy? Even though all therapeutic processes used in CBT (including processes of behavioral provenances, such as exposure) are supposed to ultimately result in cognitive change (Alford and Beck 1997), the process designed to do it most directly is *cognitive restructuring* (Padesky 1994). In cognitive restructuring patients are encouraged to notice, assess, and modify their maladaptive cognitive patterns. Manuals for CBT therapists contain descriptions of a plethora of specific techniques used for cognitive restructuring (e.g., Leahy 2017). For example, consider a depressed client who thinks that nobody cares about her. As it turns out, in this specific situation, the recurrent thought has been triggered by the fact that her best friend forgot about her birthday. In such a case, the therapist might use *Socratic questioning* to help the client evaluate the whole situation. What is the evidence that supports the belief that nobody cares about the client? What are some possible explanations for the fact that the friend forgot about the client's birthday other than that the friend does not care about her? Is the client's conclusion based on an overgeneralization, i.e., drawing over-arching conclusions from a limited number of events? By answering these questions, the client can achieve a more realistic outlook on the situation that triggered her distress, and, possibly, on her relationship with others more generally. Crucially, the cognitive practice — if successful — will likely lead to an emotional reappraisal of the whole situation and change the client's affective state. For example, the client might no longer feel neglected or ignored by her friend (even if she remains sad or disappointed that the friend forgot about her birthday).[2] The list of similar cognitive restructuring techniques is long and contains hypothesis testing, positivity reorientation, analysis of consequences, and more (Clark 2013; Newman 2015; Beck 2020).

Before we move on, it is important to highlight that, despite its popularity and influence, CBT's assumptions regarding both the source of mental suffering and the best ways to alleviate it are highly contentious (for some recent, critical, philosophical appraisals see e.g., Gipps 2013; Leder 2017; McEachrane 2009; Ratnayake 2022; Whiting 2006; Grodniewicz 2024). For example, approaching the problem from the perspective of the philosophy of mind, many authors point out that CBT's assumptions about the nature, organization, and the degree of control we can exercise over our mental states are naïve. In particular, even though it makes strong

claims regarding the need to revise one's maladaptive beliefs in psychotherapy, CBT seems to lack a clear way of spelling out the distinction between cognitive/doxastic content (the beliefs that need to be revised) and emotional/affective state (hypothesized to undergo modification as a result of cognitive change). Moreover, theoretical assumptions of CBT raise important ethical and social concerns (for discussion see e.g., Kingdon et al. 2017; Ratnayake and Poppe 2021). To mention just one commonly recognized issue, many people worry that CBT is founded on an unduly individualistic outlook on human nature, according to which one's mental suffering is to be dealt with individually and internally as opposed to in the context of broader social support, or even, where necessary, social change.

Clearly, many of these problems will be reiterated, if not amplified, in the context of chatbot-delivered CBT. Thus, the current paper should not be read as an uncritical endorsement of CBT or CBT-based therapeutic chatbots. Rather, our goal is to provide a conceptual framework that will hopefully prove useful to both supporters and critics of CBT and its implementation in chatbots, and as such, contribute to further productive debate.

This cursory introduction enables us to see in what way producers of mental health chatbots attempt to implement CBT. Let us take a look at an exemplary conversation with Woebot (2023; we used version 4.8.1. (214) of the app).

Woebot: Can you say more about what's going on?
User: I'm feeling nervous.
Woebot: If I've understood you correctly, this sounds like an issue with stress or anxiety, is that right?
User: [picking one of the available continuations] Yes, you got it.
Woebot: I'm sorry to hear you're dealing with this [user's name]. That can be difficult.
User: [picking the only available continuation] Yeah….
Woebot: Let's start by challenging the thoughts that are creating this feeling.

As we see in this exemplary conversation, as well as in most conversations with Woebot, the go-to therapeutic strategy is cognitive restructuring. If the user decides to continue this work, Woebot asks them to input three thoughts that are particularly troubling for the user and then helps the user to analyze them one by one. For example, if one of the thoughts is "I'm a failure," Woebot will ask whether the thought is based on any cognitive distortion such as catastrophizing, future-predicting, mind-reading (i.e., assuming we know what others are thinking), black-or-white thinking, overgeneralization, etc. At each stage, Woebot provides psychoeducation and explains the terms it uses. Moreover,

---

[2] In words of Padesky and Beck "[a]lthough cognitions are not always causally linked to emotional or behavioral disorders, cognitive theory proposes that cognitions mediate all change efforts." (Padesky and Beck 2003, p. 218). For the philosophical appraisal of this assumption see, e.g., Lacewing (2004) and McEachrane (2009).

along the way, it normalizes the user's experience by saying things like "Sadness can make the smallest tasks seem like unclimbable mountains," or "It's very easy to slip into [mind-reading] without even realizing it." Finally, it also offers tools targeting the user's affective state more directly, such as various forms of mindfulness or breathing exercises.

These are the main ways in which mental health CAIs implement the principles of CBT. CBT assumes that cognitive change mediates affective change and leads to positive therapeutic effects, and mental health chatbots help their clients achieve it. They do it by prompting clients to notice, assess, and revise their maladaptive thinking patterns while simultaneously assisting them in keeping their emotions on the level of activation conducive to cognitive work.

Arguably, they do not perform this task in the same way a human therapist would. In human-delivered psychotherapy, therapeutic techniques, and exercises happen within the frame of a human-human relationship characterized by "warmth, accurate empathy, and genuineness" (Beck 1979, p. 45). Moreover, therapists are able to react to whatever their clients *say* and whatever their clients *do* much more flexibly than chatbots. And yet, it seems that at least some valuable therapeutic work can occur in the interaction with a mental health chatbot. A growing body of evidence shows that chatbots are effective in alleviating general and specific anxiety symptoms as well as depressive states, preventing distress, and reducing substance use, thus improving mental health and the quality of life of their users (Gaffney et al. 2019; He et al. 2022, 2023; Laranjo et al. 2018; Prochaska et al. 2021; Vaidyam et al. 2019).[3] So, what are mental health chatbots and how should we characterize their interaction with human users? In the remainder of the paper, we argue that mental health chatbots are cognitive-affective artifacts.

3 Cognitive and Affective Artifacts

We use timetables and shopping lists to aid our memory, checklists to support decision-making processes, maps, and compasses to facilitate spatial navigation, and mathematical symbols and diagrams to support computations and inferences. These are just a few examples of objects developed or modified by human beings, that enhance our cognition and constitute new modes of thinking. To give all such objects a common name, Norman (1991) coined the term "cognitive artifacts."

In Norman's (1991) original definition, cognitive artifacts are "those artificial devices that maintain, display, or operate upon information in order to serve a representational function and that affect human cognitive performance" (p. 17). According to Norman, we can look at the use of cognitive artifacts from at least two perspectives. From the *system point of view*, artifacts enhance human cognitive abilities. For example, a system consisting of only a person, has a worse memory than a hybrid system consisting of a person and a notebook; worse spatial navigation ability than a system consisting of a person and a map or a compass; worse calculating ability than a system consisting of a person and an abacus or calculator, worse inferential ability than a system consisting of a person and a diagram, etc. On the other hand, from the *personal point of view*, the main contribution of cognitive artifacts is that they "change the nature of the task" (Norman 1991, p. 19) their user has to perform. The reason why a person equipped with a notebook "memorizes" better than a person who does not have access to a notebook is not that the capacity of the former person's memory changes. Instead, from the user's point of view, the task of memorizing has been replaced with a set of tasks which are less cognitively taxing, e.g.: "(1) The construction of the list; (2) Remembering to consult the list; (3) Reading and interpreting the items on the list" (Norman 1991, p. 21). Importantly, the new tasks are jointly leading to the achievement of the same goal as the original task of memorizing without the use of an artifact, i.e., remembering a set of information.

Since Norman's seminal work, several authors offered their own definitions of cognitive artifacts. Some of them focused more on the aspects related to Norman's system view, e.g., "cognitive artifacts are physical objects made by humans for the purpose of aiding, enhancing, or improving cognition" (Hutchins 1999, p. 126); while others — on the personal view and the perspective of an artifact's user, e.g., "[cognitive artifacts] contribute functionally to the performance of a cognitive task" (Heersmink 2013, p. 46).[4] Given that, in the case of therapeutic CAIs, the main question will

---

[3] Simultaneously it is worth keeping in mind that several problems associated with evaluating the effectiveness of mental health chatbots have been pointed out in these and other papers. In particular, due to rare and/or short follow-ups, long-term persistence and fade-out of the positive effects of chatbot-use are largely unknown. Moreover, while mental health chatbots are described to be driven by evidence-based therapeutic approaches, even in the case of interventions that have proven to be effective, "it is difficult to ascertain what the *active* ingredients of the interventions are" (Gaffney et al. 2019, p. 8). Thus, as He and colleagues (2022) recently pointed out, "Even though CBT is a highly structured therapy, translating a typical CBT-based psychotherapy into a chatbot setting is difficult" (p. 14).

[4] Recently, Fasoli (2018a) offered yet another definition, which brings the aspect of artifacts' affecting specific cognitive processes (Norman's system view) and the problem of modifying cognitive tasks (Norman's personal view) together and thus abandons Norman's original distinction. According to him, cognitive artifacts are:

"… physical objects that have been created or modified to contribute to the completion of a cognitive task, providing us with representations that we employ for substituting, constituting or complementing our cognitive processes, thus modifying the original cognitive task or creating a new one" (Fasoli 2018a, p. 11).

not be what are the cognitive abilities of the system consisting of a user and a chatbot but how chatbots can contribute to whatever change occurs *for* the user and *in* the user, in the rest of our discussion we will focus mainly on the personal view on cognitive artifacts.

More recently, by analogy to the notion of cognitive artifacts, Giulia Piredda introduced the notion of *affective artifact*, i.e., "material or non-material objects that have the capacity to alter the affective condition of the agent" (Piredda 2020, p. 550; see also Heersmink 2021, 2022; Viola 2021, 2022). Some of the examples of affective artifacts are photographs, clothing, toys, pieces of jewelry, or even pieces of music and, at least some, linguistic expressions. Each of them can, and regularly do induce a change in the affective condition of an agent. One can feel a mixture of sadness and gratitude when looking at a picture of a now-deceased grandmother, excitement when touching an engagement ring on one's finger, or feel energized when listening to their "Workout like a Spartan!" playlist.[5]

Crucially, the categories of cognitive and affective artifacts are not mutually exclusive. As noted by Heersmink (2021) "cognitive and affective artifacts can coexist in one artifact" (p. 584). Think again about a photograph of a loved one. It can evoke a wide spectrum of emotions meeting the condition for an affective artifact. But, at the same time, it may serve as a "tool for remembering" how the person looked, in general, or on a particular occasion. Thus, it is simultaneously a cognitive artifact supporting the task of remembering. Similarly with personalization algorithms. In principle, they are designed to support a cognitive task of searching for the content we might find interesting and adequate to our needs. At the same time, they have the capacity to alter our emotional states, which is often weaponized by big tech companies aware that, in our online behavior, anger and frustration equals engagement.

The goal of the next section is to argue that we can also think about computer programs — and specifically chatbots — as cognitive and affective artifacts.

4 Chatbots as Cognitive and Affective Artifacts

It is worth keeping in mind that while cognitive artifacts can be as simple as a string wrapped around one's finger serving as a reminder that one has to make ice for a party, they can also be as complicated as computer programs, applications, or even machine learning algorithms and large language models (LLMs) (see, e.g., Brey 2005, 2008; Visser 2006; Fasoli 2018b).

One of the first to explicitly recognize the special status of computer programs among other cognitive artifacts, was Brey (2005). He points out that most computer programs not only assist users in achieving cognitive goals by generating new, less cognitively taxing tasks leading to their achievement but, crucially, they often do it by creating a simulated "world" or "environment" where the tasks take place. Many of such simulated environments resemble real-world environments in which similar cognitive tasks are typically performed. For example, one of the most important breakthroughs in personal computing happened in the early 70s when the graphical user interfaces of computer operating systems started implementing the desktop metaphor, or simply, simulating desktops. On a simulated desktop, just like on the physical ones, users can move around their files, organize them in folders, and throw them away by placing them in a simulated bin. While some tasks remain the same (e.g., opening a folder, pulling out and editing a particular file), others get modified. For example, the task of searching for a specific piece of information on a physical desktop would require physically browsing through piles of folders but in contemporary desktop-simulating software the task is replaced with a set of less cognitively taxing and time-consuming tasks, such as: (1) opening a search engine, (2) typing in the phrase or keyword you are searching for, and (3) clicking on the icon representing the appropriate file.

At least some computer programs also fall into the category of affective artifacts. The clearest example here is video games, the best of which can evoke a broad spectrum of affective states, taking a player on an emotional journey comparable with reading a captivating novel or being immersed in an entrancing movie. Many of them do it by simulating not only objects but whole worlds inhabited by simulated people. At the same time "educational games," such as games developed as evidence-based interventions for children with dyscalculia (e.g., Wilson et al. 2006), are clear examples of cognitive *and* affective artifacts. They

---

[5] Even though, in this paper we will stick to Piredda's (2020) definition of affective artifacts, we would like to point out that without further specifying the meaning of "having the capacity to alter," it seems too permissive. In particular, if having the capacity to alter one's affective condition means that there is a possible scenario in which the object alters one's affective condition, arguably all objects end up being affective artifacts. Imagine that my coworker left an empty candy bar wrapper on my desk. Seeing it I get angry, which clearly counts as an alteration of my affective state. Does this mean that an empty candy bar wrapper is an affective artifact? Without trying to offer a better definition of affective artifacts — a task which significantly exceeds the scope of the current paper — in our discussion we will focus on a subset of affective artifacts which alter one's affective condition in a, more or less, regular, predictable, and goal-oriented manner.

support the cognitive task of learning by creating an engaging and exciting experience, which increases students' motivation and eagerness to learn.

Computer programs central to our current discussion are dialogical systems such as modern-day chatbots. It is worth keeping in mind that they can be powered by very different technologies, from simple decision trees based on a series of if-then rules (e.g., *if* the product has been bought not earlier than 30 days ago *then* ask whether it has been bought in-store or online) to Large Language Models (LLMs) such as GPT-4 (produced by OpenAI-Microsoft) and BARD (Google), which are based on machine learning algorithms, trained on dozens of terabytes of text data and use billions of parameters to model predictions and decisions. As a result, while rule-based chatbots can only simulate conversations on a particular topic (such as a company's return policy), thanks to the vast amount of training data, LLMs can process and generate text on virtually any subject. This flexibility is the main reason why interactions with LLMs, at least at times, feels almost like a conversation with another person (Floridi 2023).

All currently available mental health chatbots, such as Woebot, Wysa, Youper or Tess, are specific-purpose, rule-based dialogue systems (but see Miner et al. (2016)). Technical details of these mental health chatbots are proprietary, but in general, it is safe to assume that the conversational flows used in all of them are determined by fixed rules and follow a tree-based structure. For example, users of Woebot interact with the application mainly via a limited set of pre-determined response buttons, with Natural Language Processing (NLP) algorithms allowing the chatbot to process natural language inputs as answers to only some specific questions (Darcy et al. 2022, p. 290). When it comes to the type of dialogue management, currently available mental health chatbots, including Woebot, are mostly frame-based systems (Pandey et al. 2022). In contrast with simpler finite state systems, which require following a fixed sequence of moves, frame-based systems maintain a record of the necessary information and structure their subsequent questions accordingly (Laranjo et al. 2018; Harms et al. 2019). Incorporating NLP algorithms and frame-based dialogue management into chatbots aims to make the interaction with them more human-like and thus simulate a real conversation more effectively. Nevertheless, the interactions still remain restricted to narrowly-delineated topics, and aim to deal with only a few well-specified tasks (Laranjo et al. 2018).[6]

Crucially, instead of simulating objects, e.g., a desktop, chatbots simulate people, or more specifically, a social situation — a conversation with another person.[7] For example, a customer service chatbot can be characterized as a cognitive artifact, which helps its user acquire information on whether they can return a product they bought by transforming the task of reading a multipage document describing the return policies of a company into a set of tasks consisting of engaging in a simulated conversation, which involves answering simple questions such as "When has the purchase been made?," "Has the product been bought in-store or online?," "Has the product been used?"

Most of the existing chatbots are designed as cognitive artifacts which help us search for information, compare, and classify available options, etc. However, some chatbots can be equally well characterized as affective artifacts. A great example are chatbots designed to support their users' grieving practices (e.g., Elder 2020; Krueger and Osler 2022). Even though grief is a multidimensional experience with complicated phenomenology (e.g., Fuchs 2018; Mehmel 2023), it seems uncontroversial that at least one of its components is affective. As pointed out by Krueger and Osler (2022), the affective component is precisely what differentiates grief from mourning — the latter having a more ritualistic character and not necessarily being accompanied by an emotional experience. Thus, grief is "a selective emotional process that we actively regulate in diverse ways" (Krueger and Osler 2022, p. 227). At least some of these ways involve the use of artifacts: from simple ones, like letters written by the person whose death we grieve, or tombs and other objects commemorating the deceased, to chatbots, like the one built by Eugenia Kuyda after the death of her friend, Roman Mazarenko, which was trained on the corpus of Mazaernko's text messages and gave its users the impression of conversing with Mazarenko himself. As pointed out by Krueger and Osler, chatbots like this one can support the process of grieving. Specifically, they help us establish continuing bonds with the dead by preserving what Krueger and Osler call "habits of intimacy," such as "(1) conversational practices, (2) emotion regulation, and (3) shared time" (2022, p. 233). Notably, Krueger and Osler point out

---

[6] Especially in the case of psychotherapy, the problem is unlikely to be solved by simply creating more complex and detailed conversation trees and frame-based structures. As we have argued extensively elsewhere (Grodniewicz and Hohol 2023a), one of the distinctive features of psychotherapeutic conversations, is that it is virtually impossible to predetermine the set of all relevant topics and ways in which one may want to characterize their mental suffering. That is

why, even if human therapists often support themselves with various diagnostic and psychometric tools, which help them categorize the problems they hear about, their crucial ability is to process, in a productive, helpful, and flexible way, *whatever* the clients speak about.

[7] According to the influential Media Equation theory (Reeves and Nass 1996), human beings reveal a tendency to automatically treat computers as social entities. This includes interacting with artificial systems in ways similar to how we would socially interact with flesh-and-blood persons, e.g., by assigning them individual personality traits. Moreover, people feel more comfortable when using systems, including CAIs, which simulate human characteristics to a greater extent (Ahmad et al. 2022).

that conversational practices can also have "epistemic significance" in that they "might help individuals expand their understanding of different experiences or to think through different possibilities" (p. 237). In this sense, we may say that grief-assisting chatbots are *simultaneously* cognitive and affective artifacts. As we will argue in the next section, the same is the case for therapeutic chatbots.

Before we move on, we would like to briefly address one more issue. Why do we say that chatbots "simulate conversations" instead of simply "making conversations" or "conversing"? There can be different ways of spelling out what makes a linguistic exchange a conversation, but for the purpose of the current discussion, we are inclined to accept the following simple assumption: participating in a conversation requires the ability to produce meaningful utterances or simply, to mean things. However, as recently convincingly argued by Mallory (2023), by the standards of major meta-semantic theories, chatbots cannot produce meaningful utterances (see also Mitchell and Krakauer 2023). Firstly, chatbots (at least the ones currently available) lack communicative intentions, so their outputs are not meaningful according to intentionalist standards. Secondly, assuming that following social conventions requires the ability to opt out or do otherwise, chatbots do not follow social conventions and thus fail to satisfy conventionalist standards for meaning. Finally, chatbots utterances do not "refer to objects in the world" by externalist standards because "they either don't stand in the appropriate causal chains or social relations to initial tokenings, or because the machine's linguistic 'knowledge' was not acquired by the appropriate means" (Mallory 2023, p. 108).[8]

These assessments led Mallory to propose an alternative way of making sense of the human-chatbot interactions:

Table 1 Therapeutic interventions implemented in chatbots together with respective cognitive tasks, which the interventions support, and affective impacts, which the interventions induce

Intervention	Cognitive task	Affective Impact
Mood tracking and self-monitoring	Remembering, Quantitative interpretation	Eliciting one's emotions, Examining one's emotions
Identifying one's cognitive distortions and feelings, Psychoeducation	Qualitative interpretation, Acquiring new information	Recognition, regulation, and reappraisal of emotions
Cognitive restructuring techniques (e.g., Socratic questioning, hypothesis testing)	Self-examination, Self-understanding	Leading to the most profound emotional changes
Breathing and mindfulness exercises	[*Indirect impact on all the cognitive tasks listed above*]	Emotion regulation

fictionalism about chatbots (see also Krueger and Osler 2022). On this account, while the outputs produced by chatbots are *literally* meaningless they are *fictionally* meaningful, i.e., they are meaningful within a context of a certain game of make-believe. Just as we can play make-believe by imagining that sticks are swords, pots are helmets and we are knights trying to save a princess, we can make-believe that we are discussing the return of our sweater with a shop-assistant, or our mental health struggles with a therapist or a friend, while interacting with a chatbot.[9]

5 Therapeutic Chatbots as Cognitive-Affective Artifacts

As we pointed out in Sect. 2, according to CBT — the therapeutic tradition underlying all the most popular mental health chatbots — the affective change and the overall therapeutic effect are mediated by the change in one's patterns and contents of thinking and information processing, i.e., the cognitive change (cf. Lorenzo-Luaces et al. 2015). Thus, while the positive therapeutic effect (identified, e.g., with symptom reduction or the improvements of a patient's overall functioning, see Becker et al. 2011) is the ultimate goal of therapy, CBT assumes that it results from obtaining proximal goals of modifying one's patterns of thinking and information processing and the accompanying affective state. As we said in Sect. 3, cognitive artifacts are artificial objects which contribute to the achievement of a cognitive goal, while affective artifacts are objects that have capacity to alter one's affective condition. In this section, we will enumerate ways in which mental health chatbots contribute to the achievement of the proximal goals of CBT, i.e., cognitive change and affective change, thus meeting the conditions for being cognitive and affective artifacts (Table 1).

Notably, there are other material and non-material objects, such as self-help books and recordings, which support therapeutic work in similar ways and thus can also be classified as cognitive-affective (therapeutic) artifacts. We do not focus on them in the present paper since, unlike chatbots, no one seems to be confused about whether a book or a recording is an artifact or an agent. But what, then, is so unique about therapeutic chatbots?

The main difference lies in the flexibility of user interaction, which chatbots afford. While we have to pick the right recording and open a book on a chapter addressing

---

[8] For further discussion about the possibility of extending the externalists framework to account for the content produced by Artificial

Intelligence see Cappelen and Dever (2021).

[9] Mallory (2023) himself uses the example of ELIZA, a famous chatbot designed by Weizenbaum (1966) to mimic Rogerian psychotherapy and notices that in this context "DOCTOR is a fictional character within a game of make-believe therapy" (p. 11). Nevertheless, he does not discuss therapeutic chatbots in detail.

our specific problem ourselves, chatbots are designed to recognize challenges we are facing based on natural language input and recommend an adequate intervention. This may create an impression of agency — it is as if the chatbot "knows" what the problem is and "does" what we need. Moreover, as discussed above, the therapeutic work done with chatbots happens within the context of a simulated (quasi-)therapeutic conversation, which may positively influence users' engagement while simultaneously boosting the impression of agency even further. The combination of these two factors may be responsible for the superiority of chatbot-based intervention over the use of self-help books, as indicated in some preliminary research (He et al. 2022). However, it also opens the way to new ethical challenges (Coghlan et al. 2023; Laacke 2023; Sedlakova and Trachsel 2023), the most pressing of which is the risk of tricking users into thinking that they talk with another person (Tekin 2023; Grodniewicz and Hohol 2023a).

We will now move to the discussion of specific cognitive tasks that chatbots support and the affective impact that they exert.

5.1 Mood Tracking and Self-Monitoring

Maybe the simplest functionalities implemented in mental health chatbots is mood tracking and self-monitoring. While interacting with chatbots, users can record their thoughts, feelings, daily events, and activities. In this way, the interaction with a chatbot replaces the cognitive task of remembering how one felt or what one did and thought on a given day with the task of answering simple prompts (e.g., "How are you feeling today?") and, when necessary, consulting one's previous answers in the app. Moreover, in contrast with using, e.g., pen and paper, the task of remembering to record one's thoughts, activities, and feelings gets replaced with the task of reacting to notifications/reminders appearing on a screen. In this way, the tasks leading to the goal of tracking therapeutically relevant aspects of one's life become much less cognitively taxing and demanding.

However, memory is not the only cognitive process enhanced by these simple functionalities. Another one is interpretation (Brey 2005, p. 386). Brey distinguishes two types of interpretation: quantitative (i.e., measurement) and qualitative (i.e., classification). Therapeutic chatbots can contribute to the overall goal of achieving cognitive change by supporting both. Let's focus on the quantitative interpretation first. A subjective measurement of, e.g., the intensity of one's feeling or believability of one's thoughts is a commonly used CBT technique. For example, a therapist can ask: "How much do you feel upset, and how strongly do you hold your belief? Rate your feeling [emotion] from 0–100%, where 0% corresponds to having none of that

feeling and 100% corresponds to the most intense experience of that feeling" (Leahy 2017, p. 25). Just as a therapist can support a patient in a cognitive task of assessing the intensity of their emotion or believability of their thought, so can a cognitive artifact — a mental health chatbot. In both cases, the measurement can improve a patient's/users's perspective on their mental life, and thus contribute to the desirable cognitive change.

At the same time, we cannot overlook that these cognitive tasks have a direct impact on users' affective states. Writing and reading one's self-monitoring notes and tracking one's emotional state throughout different life events elicits and helps examine our emotions. For example, one may feel relieved to notice that their mood was getting consistently better over the last two weeks. Simultaneously, one may feel reassured by the observation that an activity that used to trigger their fear response whose intensity they assessed at 8 out of 10, now triggers the response they asses at 6 or lower. Finally, and maybe most importantly, the very exposure to or "being with" the emotions we journal about, notice, and observe, is often assumed to make them more manageable and less overwhelming (e.g., Grecucci et al. 2015).

5.2 Identifying Cognitive Distortions and Feelings

As we suggested, chatbots can support not only quantitative but also qualitative interpretation (i.e., classification). The first way in which they do so is by helping users realize not only how strong their feelings are but also *what exactly* they feel. Not unlike in the case of using color charts to aid the identification of colors by changing the task of answering the question "Which color is that?" into a simpler question "Are these two colors the same?" (Brey 2005, p. 387), chatbots change the task of answering an open-ended and slightly overwhelming question "How do you feel?" or "What emotion do you experience?" with a task of choosing from a list of feelings and emotions provided in the app. For example, upon choosing the option "Track and journal," Woebot displays the question "How have you been feeling today?" together with a list of possible answers such as: lonely, anxious, sad, happy, angry, guilty, bored, etc. If the user does not find the appropriate emotion on the initial list, they can ask the chatbot to show them even more options, one of which may turn out to be an appropriate description of their current emotional state. Interpreting how one feels is not an easy task. In fact, many therapists use emotion charts, with more basic descriptions, such as fearful, angry, or surprised, in the center and more nuanced descriptions, such as helpless, ridiculed, or disillusioned, at the edges, during sessions, to make this task easier for their patients. Once again, as far as relatively fine-grained tracking and identifying one's feelings and emotional reactions can be conducive

to the overall cognitive change occurring within a patient, mental health chatbots are cognitive artifacts contributing to the achievement of this goal.

Another example of qualitative interpretation supported by mental health chatbots is the identification of cognitive distortions. In Sect. 2, we mentioned that Woebot asks whether a given thought contributing to a user's low mood is an example of catastrophizing, future-predicting, mind-reading, black-or-white thinking, overgeneralization, etc. (providing a brief definition of each of these distortions upon request). Thus again, it replaces the task of answering an open-ended and potentially overwhelming question: "What's wrong with this way of thinking?" with a set of simpler tasks, i.e., answering a series of yes-no questions: "Is this an x?"

Just as in the case of self-monitoring, mood tracking, and quantitative interpretation, qualitative interpretation of the elements of users' mental struggle has a significant impact on their affective state. Choosing among the categories suggested by the chatbot, users can recognize and name, often for the first time, what they feel and experience. Such a recognition is the first step towards further regulation or even reappraisal of emotions (Grecucci et al. 2015). The emotional reactions that we can name and categorize may be less overwhelming and paralyzing. In some cases, they may even induce the "feeling of familiarity" — not unlike our irritating acquaintances whom, even if we would prefer not to see right now, we know we are capable of bearing (Hayes et al. 1999).

5.3 Cognitive Restructuring, Self-Examination, and Self-Understanding

Finally, mental health chatbots can also support their users in achieving the goal of cognitive change by guiding them through the process of self-examination and self-reflection.[10] A strategy of Socratic questioning incorporated in mental health chatbots aims to increase users' awareness and support them in evaluating their own thoughts. Prompted by accurate and well-timed questions from a chatbot, users can examine different hypotheses they embrace about themselves, others, and the world ("the cognitive triad," Beck 1979) and analyze the consequences of their actions and inactions. At least in some cases, this may lead to deepening their self-understanding (Grodniewicz and Hohol 2023b; cf. Sedlakova and Trachsel 2023, p. 8).

The kind of self-understanding one can achieve in a psychotherapeutic process is not easy to define, but it is fairly safe to accept Lacewing's (2014) assumption that it "involves grasping the connections between one's emotions,

motivations, thoughts, and behavior, past and present, including one's interpretations of and relations with others" (pp. 154–155).[11] Arguably, therapeutic CAIs can, at least to some extent, facilitate the process of "grasping" these connections, and, in consequence, the acquisition of self-understanding of their users. They do it by supporting the users in performing cognitive tasks of forming new true beliefs (and rejecting false ones), creating new connections between available pieces of information (and, if necessary, removing the old ones), and enhancing the users' capacity for grasping, e.g., by creating a conducive context thanks to intertwining cognitive restructuring with techniques targeted towards emotion regulation such as breathing exercises and mindfulness practices (Grodniewicz and Hohol 2023b, cf. Gordon 2017). Given that interactions with a therapeutic CAI happen within the context of a simulated conversation (Brey 2005; Mallory 2023), from the users point of view, it is "as if" a therapist or "a mental health ally" (Woebot Health 2023) asked them questions designed to test their assumptions and hypotheses, and inspect their beliefs. Just as in the case of traditional psychotherapy, "grasping" is ultimately done by the clients/patients themselves, but CBT techniques delivered by the mental health chatbot make the clients/patients more capable of achieving this goal.

Importantly, the kind of self-understanding one may acquire through psychotherapy is not a mere intellectual or cognitive achievement. It is not enough to figure out what is the problem or even what are the components and causes of one's psychological struggle. A profound therapeutic self-understanding typically leads to, or even equals, a far-reaching behavioral, motivational, and affective change. Thus, as far as they support the processes of self-examination and self-reflection, therapeutic chatbots prove, once again, to be both affective and cognitive artifacts.

5.4 Breathing and Mindfulness Exercises

The last important feature of mental health chatbots we will discuss is the delivery of non-cognitive interventions and techniques, such as breathing and mindfulness exercises. For example, an app may display a bubble that expands and shrinks in the rhythm, which users follow with their inhales and exhales. Similarly, many apps are equipped with pre-recorded meditations, which users can follow to regulate their emotions and become more mindful.

---

[10] Interestingly, similar use of CAI has been suggested by Lara and Dackers (2020) in the context of moral enhancement.

[11] In Grodniewicz and Hohol (2023b) we argued that self-understanding characterized in such a way is a type of what contemporary epistemologists call "objectual understanding" (Kvanvig 2003; Grimm 2021). Objectual understanding is the type of understanding one has about a given subject matter (in this case oneself) in virtue of grasping a set of information about this subject matter (in this case information about one's emotions, motivations, thoughts, and behaviors).

Such techniques are examples of what Heersmink (2021, p. 577) calls "affective techniques." As recognized by Piredda, they often are successfully implemented in or supported by material or non-material objects:

"…if I choose a meditative piece of music, it may well have a calming effect on an agitated state of mind. In this sense, our affective relation to objects helps us manage, manipulate and, in a sense, train our affective capacities." (Piredda 2020, 561)

But it would be a mistake to think about these interventions as related exclusively to the affective impact of mental health chatbots. It is an assumption common in many therapeutic approaches that valuable cognitive work cannot be done if someone is outside their window of emotional tolerance: either hyper-aroused — experiencing panic, overwhelming fear, hypervigilance, etc., or hypo-aroused — in the state of numbness or detachment. Thus, all the emotion regulation techniques have a crucial role in enabling the performance of cognitive tasks central to therapy, which, once again, reveals the status of mental health chatbots as cognitive-affective artifacts.

At this point, the reader may have an impression that therapeutic chatbots, which we discuss, are cognitive artifacts first, and that their affective impact is somehow secondary.[12] This impression is correct. However, this is not the feature of chatbots as such — after all, chatbots designed to support grieving practices are primarily affective artifacts — but of the fact that existing mental health chatbots implement CBT, whose core assumption is that cognitive change triggers affective change, and not the other way round. The cognitive change, in turn, is brought about by a set of techniques and therapeutic interventions (cognitive tasks), which, at least according to the designers of existing therapeutic chatbots, can be relatively easily implemented in interactions with artificial conversational systems. Moreover, a strength of CBT — at least from the chatbot-design perspective — is that, while it recognizes the importance of therapeutic relationship (Beck 1979; cf. Leahy 2008), it does not conceptualize it as the main mediator of therapeutic change. Therefore, much more than it would be the case in, e.g., humanistic, or existential psychotherapies, developers can try to build chatbots delivering what works in CBT, even if it remains questionable whether we can have a therapeutic relationship with a chatbot (Brown and Halpern 2021; Tekin 2023; Tong et al. 2022; Vagwala and Asher 2023, Grodniewicz and Hohol 2023a). At the same time, assuming that conversational artificial intelligence systems will likely become more and more human-like in the future,

they may also become more able to exert a direct affective impact on the patients. This, in turn, may open new paths to developing chatbots based on principles of therapeutic traditions other than CBT.

6 Conclusions

The goal of this paper was to offer a theoretical scaffolding that would allow us to accurately conceptualize the nature and role played by therapeutic chatbots in mental health care. We did it by appealing to concepts of cognitive and affective artifacts. We claim that mental health chatbots are a subtype of cognitive-affective artifacts, which contribute to positive therapeutic change by (i) simulating a (quasi-)therapeutic interaction, (ii) supporting the performance of cognitive tasks, and (iii) altering the affective condition of their users. The tasks that users perform while interacting with chatbots lead to a specific cognitive change. As it is assumed by the cognitive model of psychopathology, this change mediates the affective change and, ultimately, the positive therapeutic change (measured by the reduction of symptoms or improvement of overall functioning).

This explains why virtually all existing mental health chatbots are said to implement the principles and techniques of the therapeutic tradition built around the assumptions of the cognitive model, i.e., CBT. Simultaneously, our discussion elucidates important limitations for the development and use of therapeutic chatbots. As cognitive-affective artifacts, they can contribute to the performance of certain cognitive tasks and influence users' affective condition. However, many psychotherapeutic traditions (e.g., psychoanalytic/psychodynamic, existential, humanistic, etc.) do not share the CBT's assumption that cognitive change is necessary or sufficient for positive therapeutic change. Thus, chatbots are not (and most likely will not be soon) well suited to supplement (not to mention substitute) therapies other than CBT.

It is virtually impossible to predict the shape of human-computer interactions in the context of psychotherapy and mental health care in five or ten years. Some people suggest that, within this time frame, AI systems will become nearly human-like intelligent and able to apply their intelligence flexibly to an unlimited set of tasks. Maybe then we will be able to speak about fully fledged digital therapists and consider them agents (for discussion see Grodniewicz and Hohol 2023a). Until then, however, the framework of cognitive and affective artifacts, sufficient to characterize our interactions with maps, photographs, word processing programs, and video games will also suffice to characterize the nature of interactions with therapeutic chatbots.

---

[12] An anonymous reviewer had this very impression, for which we are grateful, as it pushed us to clarify this issue.

Declarations

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

References

Abd-Alrazaq AA, Alajlani M, Alalwan AA et al (2019) An overview of the features of chatbots in mental health: a scoping review. Int J Med Inf 132:103978. https://doi.org/10.1016/j.ijmedinf.2019.103978

Abd-Alrazaq AA, Rababeh A, Alajlani M et al (2020) Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. J Med Internet Res 22:e16021. https://doi.org/10.2196/16021

Ahmad R, Siemon D, Gnewuch U, Robra-Bissantz S (2022) Designing personality-adaptive conversational agents for mental health care. Inf Syst Front 24:923–943. https://doi.org/10.1007/s10796-022-10254-9

Ahmed A, Hassan A, Aziz S et al (2023) Chatbot features for anxiety and depression: a scoping review. Health Inf J 29:146045822211467. https://doi.org/10.1177/14604582221146719

Alford BA, Beck AT (1997) The integrative power of cognitive therapy. Guilford Press, New York

Beck AT (1967) Depression: clinical, experimental, and theoretical aspects. Harper & Row, New York

Beck AT (1979) Cognitive therapy of depression. Guilford Press, New York

Beck AT, Haigh EAP (2014) Advances in cognitive theory and therapy: the generic cognitive model. Annu Rev Clin Psychol 10:1–24. https://doi.org/10.1146/annurev-clinpsy-032813-153734

Becker KD, Chorpita BF, Daleiden EL (2011) Improvement in symptoms versus functioning: how do our best treatments measure up? Adm Policy Ment Health Ment Health Serv Res 38:440–458. https://doi.org/10.1007/s10488-010-0332-x

Beck JS (2020) Cognitive behavior therapy, Third Edition: basics and Beyond, 3rd edn. Guilford Publications, New York

Boucher EM, Harake NR, Ward HE et al (2021) Artificially intelligent chatbots in digital mental health interventions: a review. Expert Rev Med Devices 18:37–49. https://doi.org/10.1080/17434440.2021.2013200

Brey P (2005) The epistemology and ontology of human-computer interaction. Minds Mach 15:383–398. https://doi.org/10.1007/s11023-005-9003-1

Brey P (2008) The computer as cognitive artifact and simulator of worlds. In: Briggle A, Waelbers K, Brey P (eds) Current issues in computing and philosophy. IOS Press, Amsterdam, pp 91–103

Brown JEH, Halpern J (2021) AI chatbots cannot replace human interactions in the pursuit of more inclusive mental healthcare. SSM - Ment Health 1:100017. https://doi.org/10.1016/j.ssmmh.2021.100017

Cappelen H, Dever J (2021) Making AI intelligible: philosophical foundations. Oxford University Press, Oxford

Clark DA (2013) Cognitive restructuring. In: Dozois D (ed) The Wiley handbook of cognitive behavioral therapy. John Wiley & Sons, Ltd, New York

Coghlan S, Leins K, Sheldrick S et al (2023) To chat or bot to chat: ethical issues with using chatbots in mental health. Digit Health 9:1–11. https://doi.org/10.1177/20552076231183542

Cook JM, Biyankova T, Elhai J et al (2010) What do psychotherapists really do in practice? An internet study of over 2,000 practitioners. Psychother Chic 47:260–267. https://doi.org/10.1037/a0019788

Darcy A, Beaudette A, Chiauzzi E et al (2022) Anatomy of a Woebot® (WB001): agent guided CBT for women with postpartum depression. Expert Rev Med Devices 19:287–301. https://doi.org/10.1080/17434440.2022.2075726

David D, Cristea I, Hofmann SG (2018) Why cognitive behavioral therapy is the current gold standard of psychotherapy. Front Psychiatry 9. https://doi.org/10.3389/fpsyt.2018.00004

Dobson KS, Dozois DJA (eds) (2010) Handbook of cognitive-behavioral therapies. Guilford Press, New York

Elder A (2020) Conversation from beyond the grave? A neo-confucian ethics of chatbots of the dead. J Appl Philos 37:73–88. https://doi.org/10.1111/japp.12369

Fasoli M (2018a) Substitutive, complementary and constitutive cognitive artifacts: developing an interaction-centered approach. Rev Philos Psychol 9:671–687. https://doi.org/10.1007/s13164-017-0363-2

Fasoli M (2018b) Super artifacts: personal devices as intrinsically multifunctional, meta-representational artifacts with a highly variable structure. Minds Mach 28:589–604. https://doi.org/10.1007/s11023-018-9476-3

Fitzpatrick KK, 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 Ment Health 4:e7785. https://doi.org/10.2196/mental.7785

Floridi L (2023) AI as Agency without Intelligence: on ChatGPT, large Language models, and other Generative models. Philos Technol 36. https://doi.org/10.1007/s13347-023-00621-y

Fuchs T (2018) Presence in absence. The ambiguous phenomenology of grief. Phenomenol Cogn Sci 17:43–63. https://doi.org/10.1007/s11097-017-9506-2

Gaffney H, Mansell W, Tai S (2019) Conversational agents in the treatment of mental health problems: mixed-method systematic review. JMIR Ment Health 6:e14166. https://doi.org/10.2196/14166

Gipps R (2013) Cognitive behavior therapy: a philosophical Appraisal. In: Fulford KWM, Davies M, Gipps R et al (eds) The Oxford Handbook of Philosophy and Psychiatry. Oxford University Press, Oxford, pp 1245–1263

Gordon EC (2017) Social epistemology and the acquisition of understanding. In: Grimm S, Baumberger C, Ammon S (eds.)

Explaining understanding: New perspectives from epistemology and philosophy of science. Taylor & Franci, New York, pp. 293–317

Grecucci A, Theuninck A, Frederickson J, Job R (2015) Mechanisms of social emotion regulation: from neuroscience to psychotherapy. In: Bryant ML (ed) Handbook of emotion regulation. Nova Publishers, New York

Grimm S (2021) Understanding. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy, Summer 2021. Metaphysics Research Lab, Stanford University

Grodniewicz JP (2024) Belief revision in psychotherapy. *Synthese*. https://doi.org/10.1007/s11229-024-04523-0

Grodniewicz JP, Hohol M (2023a) Waiting for a digital therapist: three challenges on the path to psychotherapy delivered by artificial intelligence. Front Psychiatry 14:1190084. https://doi.org/10.3389/fpsyt.2023.1190084

Grodniewicz JP, Hohol M (2023b) Therapeutic conversational artificial intelligence and the acquisition of self-understanding. American Journal of Bioethics 23:59–61. https://doi.org/10.1080/15265161.2023.2191021

Harms J-G, Kucherbaev P, Bozzon A, Houben G-J (2019) Approaches for dialog management in conversational agents. IEEE Internet Comput 23:13–22. https://doi.org/10.1109/MIC.2018.2881519

Hayes SC, Strosahl KD, Wilson KG (1999) Acceptance and Commitment Therapy: an experiential approach to behavior change. Guilford Press, New York

Heersmink R (2013) A taxonomy of cognitive artifacts: function, information, and categories. Rev Philos Psychol 4:465–481. https://doi.org/10.1007/s13164-013-0148-1

Heersmink R (2021) Varieties of artifacts: embodied, perceptual, cognitive, and affective. Top Cogn Sci 13:573–596. https://doi.org/10.1111/tops.12549

Heersmink R (2022) Human uniqueness in using tools and artifacts: flexibility, variety, complexity. Synthese 200:442. https://doi.org/10.1007/s11229-022-03892-8

He Y, Yang L, Li T et al (2023) Conversational Agent interventions for Mental Health problems: systematic review and Meta-analysis of Randomized controlled trials. J Med Internet Res 25. https://doi.org/10.2196/43862

He Y, Yang L, Zhu X et al (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:e40719. https://doi.org/10.2196/40719

Hutchins E (1999) Cognitive artifacts. In: Wilson R, Keil F (eds) The MIT encyclopedia of the cognitive sciences. The MIT Press, Cambridge, MA, pp 126–128

Kingdon D, Maguire N, Stalmeisters D, Townend M (2017) CBT values and Ethics. SAGE, Los Angeles

Krueger J, Osler L (2022) Communing with the dead online: chatbots, grief, and continuing bonds. J Conscious Stud 29:222–252. https://doi.org/10.53765/20512201.29.9.222

Kvanvig J (2003) The value of knowledge and the pursuit of understanding. Cambridge University Press, New York

Laacke S (2023) Bias and epistemic injustice in conversational AI. Am J Bioeth 23:46–48. https://doi.org/10.1080/15265161.2023.2191055

Lacewing M (2004) Emotion and cognition: recent developments and therapeutic practice. Philos Psychiatry Amp Psychol 11:175–186. https://doi.org/10.1353/ppp.2004.0054

Lacewing M (2014) Psychodynamic psychotherapy, insight, and therapeutic action. Clin Psychol Sci Pract 21:154–171. https://doi.org/10.1111/cpsp.12065

Lara F, Deckers J (2020) Artificial intelligence as a socratic assistant for moral enhancement. Neuroethics 13:275–287. https://doi.org/10.1007/s12152-019-09401-y

Laranjo L, Dunn AG, Tong HL et al (2018) Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc 25:1248–1258. https://doi.org/10.1093/jamia/ocy072

Leahy RL (2008) The therapeutic relationship in cognitive-behavioral therapy. Behav Cogn Psychother 36:769–777. https://doi.org/10.1017/S1352465808004852

Leahy RL (2017) Cognitive therapy techniques: a practitioner's guide, Second edition. The Guilford Press, New York

Leder G (2017) Know thyself? Questioning the theoretical foundations of cognitive behavioral therapy. Rev Philos Psychol 8:391–410. https://doi.org/10.1007/s13164-016-0308-1

Lorenzo-Luaces L, German RE, DeRubeis RJ (2015) It's complicated: the relation between cognitive change procedures, cognitive change, and symptom change in cognitive therapy for depression. Clin Psychol Rev 41:3–15. https://doi.org/10.1016/j.cpr.2014.12.003

Mallory F (2023) Fictionalism about chatbots. Ergo 10:1082–1100. https://doi.org/10.3998/ergo.4668

McEachrane M (2009) Capturing emotional thoughts: the philosophy of cognitive-behavioral therapy. In: Gustafsson Y, Kronqvist C, McEachrane M (eds) Emotions and understanding. Palgrave Macmillan, London, pp 91–106

Mehmel C (2023) Grief, disorientation, and futurity. Phenomenol Cogn Sci 22:991–1010. https://doi.org/10.1007/s11097-021-09752-z

Miner AS, Milstein A, Schueller S et al (2016) Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. JAMA Intern Med 176:619. https://doi.org/10.1001/jamainternmed.2016.0400

Mitchell M, Krakauer DC (2023) The debate over understanding in AI's large language models. Proc Natl Acad Sci 120:e2215907120. https://doi.org/10.1073/pnas.2215907120

Newman CF (2015) Cognitive restructuring/cognitive therapy. In: Nezu AM, Nezu CM (eds) Oxford Handbook of cognitive and behavioral therapies. Oxford University Press, Oxford, pp 118–141

Norman DA (1991) Cognitive artifacts. In: Carroll JM (ed) Designing interaction: psychology at the human-computer interface. Cambridge University Press, Cambridge, pp 17–38

O'Donohue WT, Ferguson KE (2016) Historical and philosophical dimensions of contemporary cognitive-behavioral therapy. In: Nezu CM, Nezu AM (eds) The Oxford handbook of cognitive and behavioral therapies. Oxford University Press, New York, pp 7–27

Padesky CA (1994) Schema change processes in cognitive therapy. Clin Psychol Psychother 1:267–278. https://doi.org/10.1002/cpp.5640010502

Padesky CA, Beck AT (2003) Science and philosophy: comparison of cognitive therapy and rational emotive behavior therapy. J Cogn Psychother 17:211–224. https://doi.org/10.1891/jcop.17.3.211.52536

Pandey S, Sharma S, Wazir S (2022) Mental healthcare chatbot based on natural language processing and deep learning approaches: Ted the therapist. Int J Inf Technol 14:3757–3766. https://doi.org/10.1007/s41870-022-00999-6

Piredda G (2020) What is an affective artifact? A further development in situated affectivity. Phenomenol Cogn Sci 19:549–567. https://doi.org/10.1007/s11097-019-09628-3

Prochaska JJ, Vogel EA, Chieng A et al (2021) A randomized controlled trial of a therapeutic relational agent for reducing substance misuse during the COVID-19 pandemic. Drug Alcohol Depend 227:108986. https://doi.org/10.1016/j.drugalcdep.2021.108986

Prochaska JO, Norcross JC (2018) Systems of psychotherapy: a transtheoretical analysis. Oxford University Press, Oxford

Ratnayake S (2022) It's been utility all along: an alternate understanding of cognitive behavioral therapy and the depressive realism

hypothesis. Philos Psychiatry Psychol 29:75–89. https://doi.org/10.1353/ppp.2022.0013

Ratnayake S, Poppe C (2021) Ethical issues in cognitive-behavioral therapy. In: Trachsel M, Gaab J, Biller-Andorno N et al (eds) Oxford Handbook of Psychotherapy Ethics. Oxford University Press, Oxford, pp 514–528

Reeves B, Nass CI (1996) The media equation: how people treat computers, television, and new media like real people and places. Cambridge University Press, Cambridge

Safran JD, Muran JC (2006) Has the concept of the therapeutic alliance outlived its usefulness? Psychother Theory Res Pract Train 43:286–291. https://doi.org/10.1037/0033-3204.43.3.286

Sedlakova J, Trachsel M (2023) Conversational artificial intelligence in psychotherapy: a new therapeutic tool or agent? Am J Bioeth 23:4–13. https://doi.org/10.1080/15265161.2022.2048739

Tekin Ş (2023) Ethical issues surrounding artificial intelligence technologies in mental health: psychotherapy chatbots. In: Robson GJ, Tsou JY (eds) Technology Ethics. Routledge, New York

Tong F, Lederman R, D'Alfonso S et al (2022) Digital therapeutic alliance with fully automated mental health smartphone apps: a narrative review. Front Psychiatry 13:819623. https://doi.org/10.3389/fpsyt.2022.819623

Vagwala MK, Asher R (2023) Conversational artificial intelligence and distortions of the psychotherapeutic frame: issues of boundaries, responsibility, and industry interests. Am J Bioeth 23:28–30. https://doi.org/10.1080/15265161.2023.2191050

Vaidyam AN, Wisniewski H, Halamka JD et al (2019) Chatbots and conversational agents in mental health: a review of the psychiatric landscape. Can J Psychiatry 64:456–464. https://doi.org/10.1177/0706743719828977

Viola M (2021) Three varieties of affective artifacts: feeling, evaluative and motivational artifacts. Phenomenol Mind 20:228–242. https://doi.org/10.17454/pam-2000

Viola M (2022) Seeing through the shades of situated affectivity. Sunglasses as a socio-affective artifact. Philos Psychol 1–25. https://doi.org/10.1080/09515089.2022.2118574

Visser W (2006) The cognitive artifacts of designing. Routledge, Mahwah

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

Whiting D (2006) Why treating problems in emotion may not require altering eliciting cognitions. Philos Psychiatry Psychol 13:237–246

Wilson AJ, Revkin SK, Cohen D et al (2006) An open trial assessment of the number race, an adaptive computer game for remediation of dyscalculia. Behav Brain Funct 2:20. https://doi.org/10.1186/1744-9081-2-20

Woebot Health (2023) Woebot. https://woebothealth.com

Zilcha-Mano S (2017) Is the alliance really therapeutic? Revisiting this question in light of recent methodological advances. Am Psychol 72:311–325. https://doi.org/10.1037/a0040435