Philosophical Psychology

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/cphp20

Disrupted self, therapy, and the limits of conversational AI

Dina Babushkina & Bas de Boer

To cite this article: Dina Babushkina & Bas de Boer (03 Sep 2024): Disrupted self, therapy, and the limits of conversational AI, Philosophical Psychology, DOI: 10.1080/09515089.2024.2397004

To link to this article: https://doi.org/10.1080/09515089.2024.2397004

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

Published online: 03 Sep 2024.

Submit your article to this journal

Article views: 85

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=cphp20

PHILOSOPHICAL PSYCHOLOGY
https://doi.org/10.1080/09515089.2024.2397004

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Disrupted self, therapy, and the limits of conversational AI

Dina Babushkina and Bas de Boer

Faculty of Behavioural, Management and Social Sciences, Philosophy Section, University of Twente, Enschede, The Netherlands

ABSTRACT

Conversational agents (CA) are thought to be promising for psychotherapy because they give the impression of being able to engage in conversations with human users. However, given the high risk for therapy patients who are already in a vulnerable situation, there is a need to investigate the extent to which CA are able to contribute to therapy goals and to discuss CA's limitations, especially in complex cases. In this paper, we understand psychotherapy as a way of dealing with existential situations and position CAs in the context of the therapeutic experience of patients. This experience is determined by the patient's unique personal context and specific therapy goals. We suggest that psychotherapy is a fundamentally dialogical activity, because it crucially involves work on the self and one's self-narrative. This brings us to our central question: is it possible for CAs to engage in a productive therapeutic dialogue, given their limitations as epistemic agents? We will discuss several of those limitations, show how these undermine the possibility of engaging in a therapeutic dialogue, and illustrate those limitations through discussions of the cases of grief and abuse.

ARTICLE HISTORY
Received 21 February 2024
Accepted 20 August 2024

KEYWORDS
Psychotherapy; ethics and epistemology of AI; existential situation; philosophy of grief; philosophy of abuse; conversational agents

1. Conversational agents for psychotherapy: claims and hopes

The development of digital technologies over the past decades has been accompanied with promises about how they can change or improve various aspects of the medical sector: from mobile applications enabling people to receive health-related information and advice to telecare technologies, positioned as to offer medical care from a distance. And now conversational agents (CAs) come with the promise of a positive impact on the health and mental health sector (e.g., Denecke & May, 2023; Hocking et al., 2023; Sarkar et al., 2023; Tudor Car et al., 2020; Viduani et al., 2023).

A CA is a type of artificial intelligence that engages in or supports conversational interactions through the use of text, images, voice or

CONTACT Dina Babushkina d.babushkina@utwente.nl Faculty of Behavioural, Management and Social Sciences, Philosophy Section, University of Twente, P.O. Box 217, Enschede 7500 AE, The Netherlands

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

a combination of those (e.g., Bérubé et al., 2021, p. 1; Jabir et al., 2023). The ability of CAs to exchange phrases with users has vastly increased due to the development of large language models (e.g., ChatGPT).[1] Whereas they were initially used to assist short-term transactional tasks (e.g., delivering information relevant to one's health), CAs are increasingly ascribed independent roles and positioned as relevant to long-term projects such as disease management and psychotherapeutic counseling (Abd-Alrazaq et al., 2019; Pham et al., 2022; Tudor Car et al., 2020, p. 10). Ever more often is it asked if a CA can be seen a substitute for a doctor (e.g., Bunz & Braghieri, 2022; Haemmerli et al., 2023), therapist (e.g., Carlbring et al., 2023; Fiske et al., 2019; Graber-Stiehl, 2023), or an assistant to them (e.g., Eshghie & Eshgie, 2023).

A widely discussed reason for allocating a more active and independent role to CAs is that conversations, while central to psychotherapy, are very time-consuming. Since access to mental healthcare is an issue of distributive justice, CAs are thought to be of help when addressing issues related to care allocation (e.g., Tekin, 2021). CAs, then, are thought to be able to support existing mental health infrastructures while reducing the workload of therapists. For instance, it is suggested that they can be useful to those who are currently on a waiting list for receiving appropriate care, those that are hesitant to seek care due to a fear for being stigmatized, or to those who have limited access to mental health support otherwise (e.g., Vaidyam et al., 2019, p. 458).

However, whether CAs can contribute to better care allocation (e.g., Brown and Halpern (2021) cast doubts on this claim) depends on the extent to which they can be considered a meaningful alternative to existing forms of mental health support and do not compromise the goals of therapy itself. Whereas technology-driven research that highlights the benefits of CAs (e.g., Ahmad et al., 2022) or suggest ways to overcome user resistance (e.g., Jabir et al., 2023; Prakash & Das, 2020) is plentiful, critical evaluation of the effectiveness of such claims and aspirations seems to be lagging behind (e.g., Vaidyam et al., 2019). That is, studies find that users report positive experiences about CAs in a variety of contexts (e.g., boosting mental resilience, mitigating symptoms of depression and anxiety), but little is known about the long-term effects of using CAs, nor about the underlying mechanisms that are explanatory for the positive experiences (e.g., Li et al., 2023). Moreover, it is unclear how CAs align with basic assumptions in psychotherapy such as the formation of an affectionate bond between client and therapist or the healing role of the dialogue that are central to psychotherapeutic practice (e.g., Herbener et al., 2024).

In this paper, we explicitly position CAs in relation to such basic assumptions in psychotherapy and explore if and to what extent CAs are capable of having therapeutic conversations. Given the pace at which AI technology

develops, it seems reasonable to assume that, even though most current CAs are linguistically limited, more advanced linguistic interactions can be expected in the near future (Grodniewicz & Hohol, 2023). This might feed the idea that CAs should not be considered as a mere additional support tool (comparable, for example, to self-help books), but should better be compared to what is expected from a psychotherapist (Ibid., 3). However, in that case, CAs should not only be evaluated in terms of accuracy and efficiency but also in terms of how well they realize desirable therapeutic processes and goals.

The complexity of human psychology, psychotherapeutic theory and practice (which requires not only knowledge but also personal experience of the therapeutic process) sometimes leads to comments expressing that it is unclear both what exactly psychotherapy consists in (e.g., Grodniewicz & Hohol, 2023, p. 3), and what would count as a recovery from a particular psychological problem (e.g., Meadows et al., 2020). However, while there are many definitions available and many different methodologies used in psychotherapeutic practice,[2] it is safe to say that the ultimate goal of any therapy is to help the patient to cope with a psychological problem. It is useful at this point to clarify our standpoint: we do not argue for a specific approach to therapy. Rather, we aim to discuss, from an analytical point of view, what matters for a patient *as a patient* in the therapy process[3] and, from the normative point of view, what they are justified in expecting from it.[4] On this very basic level, it is not unreasonable to expect that when offered a certain therapy option, the person experiencing the psychological problem has an interest in making sure that this option can help address the problem as she is experiencing it.

One of the appeals of integrating CAs into a therapeutic context is the idea that they can engage in a dialogue. Some even suggest that one option for a fully autonomous internet-based self-help tool would be "to integrate AI, in the form of a conversational agent, into the platforms [. . .] either as a tutor that becomes a virtual psychoeducational coach available to the client at all hours of the day, or taking on the role of a psychotherapist, giving constructive feedback and posing clinically informed inquiries such as Socratic questions that could directly help the client when analyzing negative automatic thoughts or maladaptive core beliefs" (Carlbring et al., 2023, p. 1). There are multiple similar claims in the literature, but this one most explicitly appeals to one of the most complex and advanced forms of dialogue that requires considerable investment and high-level cognitive skills. Essentially, Carlbring et al. (2023) suggest that a CA is capable of a meaningful conversation with a patient, to the effect of reaching insights and deep understanding, and thus, addressing the patient's problem by affecting in a goal-oriented, desirable manner her patterns of thinking, self-reflection, feeling and action.

This prompts the question: is a patient in fact justified in believing that CAs can meaningfully address the experienced problem? Sedlakova and Trachsel (2023) partly address this issue when suggesting that CAs should be seen and evaluated as mere tools when used to transfer information, whereas if we want them to engage in therapeutic conversations, their ability to meet standards of agency should be evaluated. They argue that the latter is doubtful: even when granting that CAs can *mimic conversations*,[5] they lack the capacities that we would expect from human agents in a conversation (e.g., social sensitivity, potential to build a relationship, exhibiting a normative stance). As a result, we ought to be critical of CAs in therapeutic context, because the term "conversational agent" is a misnomer that might raise wrong expectations. Fuchs (2024) makes a similar point by arguing that since empathy is a uniquely human capacity, therapeutic conversations cannot and should not be delegated to CAs. We do agree that CAs lack genuinely human capacities, but – and this is important – the appeal of CAs is the claim *that mimicking is enough*.[6] In the end, anything, to a degree, can be mimicked or simulated, even empathy. And so, the question that we need to ask is: To what extent are simulations of dialogical interactions conducive to therapy goals and to what extent are they acceptable? When is simulation detrimental to these goals and should be avoided? Our intention here is to delineate cases where simulation is not enough, such as complex cases connected to existential situations.

We will proceed as follows: first, we discuss therapy as a response to an existential situation, which a person has to live through and which is associated with the shattering of one's familiar relationship to one's surroundings. We suggest that this shattering disrupts one's self-narrative such that one's feeling of coherency is lost, and that restoring one's sense of self requires engaging in a dialogue, which requires a much deeper engagement with the internal structures underlying the shattered world than mere conversing can offer. By conceptualizing therapeutic dialogue as an epistemic existential tool that allows rebuilding of the self and one's relationship to others, we point to its intersubjective nature. We then show that there are cases where merely mimicking such dialogues is insufficient for meeting the therapeutic needs, and this draws a strong line in the possible applications of CAs in therapy.

2. Artificial therapists and patients' lived experience

The introduction of conversational agents into psychotherapy brings to light a clash between two competing tendencies characteristic of the concept of personalized medicine: on the one hand, this concept appeals to optimization of measuring health-relevant parameters, while on the other hand, it aims at putting patients' health interests at the centre. The

former tendency relies on the understanding of medicine in general and therapy more specifically as an institution that requires resources, falls under healthcare regulations and can be optimized by using digital technologies (for instance, through digital phenotyping). From this point of view, the goal is to make therapy cost-efficient, quick, and easy to use. At the same time, the narrative about patient-centered care rather focuses on unique circumstances of a person in the need of therapy and encourages solutions even in the situations which are not easily typified and algorithmized.

The latter starting-point highlights the patient's perspective and is also our point of departure. We are concerned with what matters for the patient *as a patient*, their health goals, their lived experiences, abilities and limitations, and the extent to which therapy can bring relief: the reasons why someone would go into psychotherapy in the first place. These reasons can be vastly different,[7] ranging from traumatic responses to events like rape or life-threating situations, to feelings of grief, to having been in an abusive situation, to not feeling at ease in one's current romantic relationship. Not only can the reasons for searching therapy be different, also the specific context of the individual patient matters. For example, the experience of having been in an abusive situation is very different if the abuser is a parent or a spouse, and if abuse is experienced as a child or an adult. Also, the experience of grief might be very different depending on one's relationship to the deceased, the type of loss that one experiences, and whether the environment has prepared you for dealing with loss in a constructive manner.

We use the general term of existential situation to refer to all of these different events and experiences. While there are obviously differences in intensity and severity, the goal of therapy in all situations is similar: namely that patients desire for relief. Introducing the term "existential" enables us to sharpen our view on CAs; an important criterion by which they can be evaluated is the extent to which they can be meaningful in helping to cope with an existential situation and alleviating the burden that this situation puts on them. To be sure, when talking about "existential situations" and "the experience of therapy", we do not refer to personal idiosyncratic perceptions of patients. While these are valuable in themselves, we need to distinguish between patients being satisfied with therapy/a therapist – amounting to "liking/disliking" – and the experiential logic. The latter can be understood as a rational evaluation of one's experience based on the experiences of trauma and idiosyncratic features of the lived situation (facts about causal connections, psychological states, neurological peculiarities etc), of whether certain method/intervention helps. Our focus in this paper is on the latter as we think it is important to ensure that the therapy situation is construed in such a way that makes room for any such

idiosyncrasies (and especially most complex and difficult cases) that is characteristic of one's own way of "mastering the trauma" (Brison, 2003, p. 54). This implies remaining on an abstract theoretical level of discussion, conceptualizing what matters for any patient.

3. Existential situations and the fabric of familiarity

Existential situations disrupt the initial structure of everyday life. They are characterized by an experience of no longer feeling at home in a world that was previously familiar (e.g., Aho, 2019; Svenaeus, 2000). Existential situations constitute an essential part of what it means to be a human: one cannot escape living through episodes in which one's picture of the world and oneself is fractured, when structures of familiarity are disrupted and when meaning is lost. Existential situations are paradoxical, in a sense. On the one hand, they make us understand that there are determinations of life that are not under our will: events like death or violence occur whether we want them or not. At the same time, existential situations reveal that the way we experience and deal with those events is determined by how we understand the world and our place in it.

Existential situations are fundamentally *experiential*—there is no grief without a grieving person experiencing certain emotive attitudes and thought patterns; there is no abuse without a victim, who is living through a scenario where the actions of the abuse unfold for her. The subjectivity, the idiosyncrasy of the person's response to these essential pan-human events are an inalienable part of the existential situation. And precisely because such situations are a necessary part of life, they cannot be pushed out of the horizon of experience and should be approached as situations to be dealt with by incorporating them into one's self-narrative. Coping with an existential situation involves being *able to (re-)integrate the disruption into a novel structure of familiarity and self-narrative*. Regaining familiarity alongside the development of a new self-narrative happens in the sphere of *intersubjectivity*, that is, in interaction(s) with others that have a position outside of the specific existential situation while at the same have a sense of existential situations generally. Within the therapy process, this implies that the therapist is someone who is essentially outside the existential experience, but is simultaneously equipped to act as a guide.

4. Self-narratives as interpretative tools

Psychotherapy is, in a sense, a guided way out of the trap of the existential situation. And as such, it is itself an existential experience, consisting of finding a way out of "being stuck" through changes in one's self-narrative as well as patterns of thought, desires, emotive reactions,

and ultimately, actions. A way to recover from a traumatic event is through constructing a new self-narrative, reiterating the past self into a new vision of oneself. In therapy, one is experiencing and addressing oneself, re-contextualizing the disrupted self from a more constructive and nurturing perspective. But this perspective, this reference point can only be provided by someone outside the existential trap: the existential experience rebuilding of the self is, essentially, an intersubjective experience. Brison (2003), for example, herself being a survivor or a violent (rape-to-kill) attack and having spent years on developing a philosophical account of the recovering self, describes the ability to reconstruct one's self-narrative in terms of one's personal narrative being validated by others, which enables one to develop a re-conceptualized and re-told first-person story of the experienced event. An essential role in this process belongs to the dialogue.[8]

An existential situation – especially one as extreme and violent as described by Brison – disrupts the ordinary, the familiarity of the world and ourselves. In most cases, our experience of the world is one in which the meaning of things around us is familiar prior to any form of reflection. Such a pre-reflective sense of familiarity makes it so that we can move through the world relatively unproblematically and are capable of carrying out actions habitually. Put differently, ordinary experience presupposes an assumption of familiarity both with the objects (inanimate as well as animate) one encounters and the subject (or self) that has a particular relation with those objects. This relational nature of familiarity is reflected in one's coherent self-narrative: the self which has its place in the world. An existential situation disrupts the coherency of a self-narrative[9] and accompanied embodied intentionality leading to a loss of (aspects of) one's pre-reflective sense of familiarity. Put colloquially: things appear to be "out of place", and ultimately, you start appearing to be out of place in the world (Brison, 2003, p. 78).

Regaining a sense of familiarity requires a transformation of one's self-narrative such that one can attach a (novel) meaning to the things around. However, given that it is impossible from within the existential situation to find a reference point for this new sense of self, the interaction with what can be called "the objective Other" (in this context, the therapist) becomes essential; someone who is a part of the referent world, i.e., the world outside the shattered self-narrative. This is a complex interaction that involves the re-negotiation of things that are important and meaningful, revealing and changing of certain patterns of thought and behavior. This interaction *enables* the use of a first-person narrative as *an interpretative tool*: a tool that helps to know what happened, how the objective events disrupted the inner world of someone, and thus, to find ways to amend the damage. As a means to focus attention, analyze,

and understand, the therapeutic dialogue is, thus, also an epistemic tool. As Brison (2003, p. 87; italics added) says: "understanding trauma ... requires one to take survivors' first-person narrative seriously as an essential *epistemological tool*".

5. Dialogue vs conversing

What is important to keep in mind, since the word *dialogue*[10] has a multiplicity of meanings, is that in the context of therapy, engaging in a dialogue is one of the essential ways of dealing with an existential situation.[11] It is a medium in which the interaction with "the objective Other" unfolds with the goal of re-building the self-narrative. One needs to talk the experience through and needs to have it heard by others, to be recognized and validated. But, given the epistemic dimension of therapy, dialogues are also a way to discover, determine and categorize (putting a name on) the experience and as such, they *are a way for a guided transformation of the self through intersubjectivity*. One aspect of therapy that CAs aim to subtitle, is the partner in the therapeutic dialogue, in other words, "the objective Other" in the relationship which is essentially intersubjective.

To be able to evaluate CAs capacity for such substitution, it is crucial to consider the difference between *conversing* (a dialogue in a weak or trivial sense) and *the dialogue proper* (a dialogue in the strong sense of a mutually invested intersubjective experience). Conversing is something we do casually. In a literate sense, to converse is to exchange minimally fitting responses to utterances of others. The standards of fit may vary but they can be quite low and formalistic to accommodate for what can be also called a "small talk". Examples of such interactions are greetings ("How are you?"/ "Good, how about you?"), simple requests for information ("When and where does event X take place?"), or routine conversations ("The dinner is delicious, don't you think so?"). All of these interactions take place against a horizon of familiarity, such that it is not needed to make explicit why the talk takes place, because the expectations are clear either from the purpose of the interaction itself or because the interaction is embedded in shared familiar routines. Linguistic interactions against such a horizon of familiarity are very frequent, can be rather extensive, of different levels of complexity, and serve various purposes (make one feel accepted, safe, appear polite, etc.). When long and complex enough, these can be referred to as dialogues in the weak sense, i.e., an exchange of utterances between two parties of a conversation that formally can be seen as mutually fitting responses (for example: "Where are you heading?"/ "Nietzsche was a great philosopher!" is not a dialogue even in a weak sense of conversing).

However important such everyday talk might be, the interactions it gives rise to do not require any substantial investment from the conversing partners such as making one's feelings or thoughts explicit, nor do they require a mutual effort toward a deeper understanding of a particular problem or a solution to it. These are, however, experiences we expect from a dialogue in *the strong sense* of the word. A dialogue proper has a connotation of a more substantial interaction which operates with meaning and is goal-oriented in the sense that parties are invested in a common topic or problem which they wish to clarify together, explain to each other or find a resolution to. A productive dialogue is a mutual cognitive and emotive journey toward a common destination – e.g., a Socratic dialogue is a journey to develop an understanding of a concept; disagreeing parties engage in a dialogue to find a resolution; friends can have a dialogue to deepen their understanding of life experiences (such as love and friendship itself). It does not mean that dialogues are only about deep existential topics, but they are engaging existential experiences as they are genuine events through which we achieve understanding, clarity, learn something new, convey messages, convince others, or achieve recognition.

And thus, having a dialogue often presupposes suspending a purely instrumental and superficial relationship with one's surroundings and making publicly explicit one's internal states (such as thoughts, values, and motivations). It also presupposes a certain cognitive and emotional investment into the common aim of the conversation. Dialogues carry with them an *expectation* that one's partner is able and willing to invest as well, be it in decoding the idiosyncratic symbolic complexity of one's psychic experiences, reaching and sharing an understanding, granting acknowledgment and recognition, relating emotionally and giving an appropriate emotive response, or thinking along to find a solution.

6. Is simulation enough?

It would be hard to substantiate the claim that CAs are capable of dialogue in the strong sense as described above. It is also doubtful that they are capable of conversing in the same sense as humans are – but that is beyond the point. CAs' appeal is not that they are capable of dialogue, but that they are capable of simulating or mimicking dialogs and that this simulation/mimicking is enough to achieve certain therapy goals. Even though powerful artificial agents such as CA are simulators by design, the act of simulating taps into the kind of problems that are not unique to AI. People also simulate: they cheat, pretend, fake, make an impression of being or feeling something they are not or do not feel. This is not to say that all simulation is necessarily wrong; some forms of simulation are acceptable and even encouraged. The problem at hand, however, is whether simulation is suited

for driving the therapeutic dialogue as a means to dealing with an existential situation.

In this respect, Brison's (2003, p. 9) observation of the dissonance in the victims' experience is insightful: she connects the feeling of "unreality" to the types of responses she was receiving from others after rape has taken place. She encountered a variety of such responses: forms of denial, attempts to find elements of the victim's fault, attempts to find the "bright side" or learn something from the experience, attempts to excuse what happened with reference to God's will, verbally diminishing the experience, or suggestions to "buck up". All such responses fail to appropriately acknowledge and validate the lived experience of the existential situation, specific to this particular survivor, and take the form of a cliché, or impersonal platitudinal statement. Such responses are not only unhelpful in making sense of one's experience and finding ways to heal, but they can actually be harmful since they create a feeling of invalidation of the traumatic and post-traumatic experiences. The result is that the survivor is not being heard becuase the conversation partner is not being engaged in listening to what the survivor is saying. The response fails to engage with and address the victim's message and testimony, substituting them with a generalized preconception of what their experience should be.

Under such conditions, a simulated response, which replaces a genuine emotional and cognitive engagement with the speaker, is a failed response and a failure in listening. In Brison's words: "[I]t can be traumatizing when people refuse to listen. In my case, each time someone failed to respond I felt as though I were alone again in the ravine, dying, screaming. And still no one could hear me. Or, worse, they heard me, but refused to help" (2003, p. 16). It is important to keep in mind that in virtue of its design, a CA never responds to *you*; it responds to a typical request as it is recognized and construed by the algorithm. Moreover, a CA's response is constructed in such a way that it consists of generic platitudinal statements about the type of situations you might (or might not) find yourself in, based on certain key words (used in the prompts) which happen to frequently co-occur with certain words in the data-set on which the CA has been trained. This means, that on top of being generic platitudinal statements, they are also limited to the range of content of the texts they were exposed to: if your case happens to differ in significant aspects from those texts, CAs do not have the instruments to deal with it. This boils down to CAs essentially not being able to deal with problems that are, so to say, open-ended: questions of the type "What is wrong with me?" (typical for a medical context, cf. Kennedy, 2021) or "Why am I sad?" – i.e., questions which require deep analysis of the personal situation (such cognitive involvement as understanding, connecting, and re-conceptualizing) in order to find a reason for what

is happening to *me*. A mere theoretical venture in possible explanations of typical situations of people with similar complaints is not going to be enough. As a result, when trying to "discuss" one's situation with the CA, one mayfeel irrelevant to the conversation and ignored.

Given the goals of the therapeutic process, a patient is morally justified to feel that a generic platitudinal response from a therapist is insufficient. This implies that the dialogue between the therapist and the patient is subject to moral as well as epistemological normativity: moral, insofar as the therapeutic dialogue happens within an interpersonal relationship, and epistemological, insofar as it is a cognitive event, with a goal to achieve understanding and reconceptualization. The ability to evaluate if a therapists' response sufficiently addresses one's question is an indicator that the dialogical interaction comes in degrees and, as such, is it essential to be able to raise one's concern about the progress of therapy process and ability to influence and change it, if needed. It is thus crucial for the patient's control over the process of healing. In the case of a dialogue between humans, it is possible for the patient to raise such concerns with the justifiable expectation that the conversation partner will respond to them in a meaningful and productive way (and if not, the patient can hold the therapist accountable).[12] In interaction between CAs and humans there is no such possibility – an algorithmically predetermined agent (CAs are also predetermined by their algorithms) lacks such epistemic flexibility.

Simulated conversation is morally problematic also because therapeutic dialogues create an asymmetric situation as one partner is coping with the existential situation whereas the other is guiding and represents a certain (epistemic, medical, institutional etc.) authority. Such asymmetry in the dialogical relationship with the therapist puts the patient into a vulnerable position, and this has to be taken into account when deciding the extent to which they should be exposed to a CA with limited epistemic and moral capacities. Because CAs merely simulate human capabilities, the possibility of them relating to the experience of a human subject on the existential level is precluded: for instance, they miss the capacity to relate to a grieving process or to the suffering endured when someone is recovering from abuse. This kind of engagement is an expression of a shared intersubjectivity that can be justifiably expected by the patient from the partner in the therapeutic dialogue. This is not something one can design, but rather emerges in interaction between entities with a certain ontological status.

In what follows, we focus on two types of existential situations – grief (Section 7) and abuse (Section 8) – and discuss the extent to which CAs are limited in helping to cope with those situations. We deliberately chose these two distinct examples to demonstrate the centrality of dialogues for psychotherapy in different contexts as well as the ways in which CAs may undermine the possibility of engaging in a dialogue.

7. The case of grief

It is an existential given that each of us will be confronted with loss: losses can range from getting unemployed to the death of one's pet, from a broken relationship to the death of one's life partner. Losses often result in the experience of grief, which can be broadly described as an emotional process that is a response to loss (Ratcliffe, 2022, p. 3). Grief is different from bereavement and mourning. Bereavement can be understood as the objective situation of losing someone or something that matters to us, whereas mourning refers to acting in accordance with social and cultural norms, expectations and practices surrounding loss (Ibid., p.4). Grief, in contrast, is an individual emotional experience in response to loss or bereavement and it does not need to coincide with the expectations and norms of mourning.[13]

Two recent accounts of grief are Goldie's (2011) narrative account and Ratcliffe's (2022) phenomenological account.[14] Both have in common (a) that grief is understood as a process and not as a singular response that is causally connected to a certain event (i.e., the constitution of grief cannot be explained by formulas like "if x, then y" or "iff x, then y"),[15] and (b) that this process involves a renegotiation of one's past in light of the experienced loss and the other way around. Both Goldie and Ratcliffe take particular characteristics of the process as more important than the event instantiating it. Goldie defines the grieving process as a temporally extended, unified pattern that "includes characteristic thoughts, judgments, feelings, imaginings, actions, expressive actions, habitual actions, and much else besides, unfolding over time, but none of which is essential at any particular time" (2011, p. 125). The pattern can be identified as grief because it has a recognizable "shape", despite it being specific to the situation of the grieving subject. Ratcliffe maintains that grief is a process that is characterized by the shattering of one's experiential world through the experience of loss and crucially involves "to reconcile the structure of one's experiential world with an explicit acknowledgment of loss" (2022, p. 8). They both indicate that grief is not only something that simply happens to us but something that presupposes certain types of actions or might even require certain work.

As an existential situation, grief experientially disrupts one's self-narrative or familiarity with the world because of the experienced loss: specific habits that one was used to carry out unproblematically (and which were structured around a certain aspect of one's life that is no longer there) can no longer be carried out as such. The active side of grief consists in re-negotiating a novel relationship with the world i.e., to reconcile the structure of one's experiential world. Coping with grief, therefore, can be characterized as the constitution of a novel meaningful self-narrative and/or familiar world in which the past is incorporated anew, such that novel future possibilities can emerge.

Grieving after a loss is acceptable and, in certain cases such as the loss or death of a partner or close friend, people can be expected to display signs of grief. At the same time, there are cultural norms about how long it is reasonable to grieve and when the griever should be able to "give the loss a place". This expectation is reflected in the distinction between typical grief and pathological grief.[16] It is not our aim here to introduce – in line with Ratcliffe (2022) any substantial distinction between the two, nor to determine whether certain grief experiences are indeed pathological. Our goal is to bring attention to fact that some characteristics of grief might explain the intensity of certain grieving processes. Ratcliffe proposes to understand non-ordinary grieving processes in terms of the presence of those resources necessary for reconfiguring of one's experiential world. In order to constitute a novel form of familiarity, grieving has to involve an explicit engagement with one's loss. This engagement often involves interpersonal and social processes by means of which both the past can be reconfigured and new possibilities for the future can emerge (Ratcliffe, 2022, p. 206). The absence of these aspects is, according to Ratcliffe, characteristic of non-ordinary grieving processes. Alongside the disruption of one's familiar interactions with the world, one's relationships with others are disrupted, leading to an inability to trust them. As a consequence, the griever becomes unreceptive to those interpersonal and social processes that contribute to comprehension and adjustment.

Now that we have briefly discussed what grief consists of and which resources are important in the grieving process, we are in a better position to evaluate to what extent CAs can be helpful in coping with grief. Recent developments suggest that they can. An exploratory study cautiously suggests that CAs could help in coping with grief, either by offering companionship in the form of a chatbot or by developing an avatar-based simulation of the deceased that the griever can interact with (e.g., Xygkou et al., 2023). In the literature, we find different responses to such technological developments: some authors see clear advantages for using CAs in the grieving process (e.g., Cholbi, 2022, p. 90), whereas others maintain that they necessarily lack the capacity of empathic understanding needed for dialogical engagement (e.g., Fuchs, 2024). Our analysis of grief indicates that CAs are bound to have limited use in the grieving process. First, since grieving crucially involves interpersonal and social engagement to reconfigure one's experiential world, therapeutic conversations with artificial agents have the potential to undermine the grieving process as they increase the risk of stimulating griever's tendency to shut off from interactions with people, thereby limiting chances of engagement and restoring meaningful relationships with others. In such cases, CAs can actually do harm. This means that, even if conversational agents might be useful with regard to some aspects of grieving, they should be able to express the limits of their

relevance and clearly indicate that they are not a proper replacement of the social and interpersonal and social relationships necessary for coping with grief.

Furthermore, grieving is experienced in myriad ways; it may be connected to different kinds of bereavement, have varying degrees of intensity, and is experienced in relation to one's particular personal history. Hence, grief, while being a universal existential situation, is not a generalizable experience. It is closely tied to the experiential world of the subject that is structured around particular phenomena such as habits, anticipations, or values. All of these are grounded in intersubjectivity, resulting in an experiential world where the feeling of connectedness remains but the entity that one was initially connected to is no longer there. Insofar as CAs rely on data collection, and even if we assume that it is possible (and desirable) to personalize the artificial agents, they are likely unable to capture the intricacies of the phenomenology of the grieving process. Such inability puts limits on CA's ability to actively contribute to coping with grief. Even if they are capable of indicating social expectations around grief, it is likely that grieving will be confused with mourning, which makes it difficult to see how they help users to establish a constructive relationship with their own past and the social expectations surrounding loss.[17]

In sum, there might be certain merits when it comes to conversational agents, especially in lighter cases. However, if we take grief to be an (a) temporally extended, and (b) highly diverse process, there are clear limits to the types of interventions CAs should be allowed to implement. One should resist the temptation to see grief as a generalizable process and CAs as one-fit-all solution. There is a risk for pathological tendencies to go unnoticed and undermine the griever's capacity to engage in interpersonal and social processes needed for acceptance and recovery.

8. The case of abuse

Abuse is another good example of the therapeutic limitations of artificial agents. It is difficult to come up with a definition of abuse that would encompass the multiplicity of its forms and situations where it occurs (e.g., physical and sexual abuse, emotional and psychological abuse,[18] t in such contexts as family relations, intimate partner violence (e.g., Nicolson, 2019), and in vulnerable groups as children (O'Hagan, 1995), elderly (Scheiderer, 2012), women (Tolman, 1992). As a result of the plasticity of violence, the experience of abuse is too varied to attach specific parameters and quantifiers to it, even more so in the cases of psychological abuse (Follingstad, 2007). However, from a philosophical point of view, we can conceptualize abuse as an undue treatment, maltreatment (McGee & Wolfe, 1991) or misuse. Minimally, it can be said to have to do with a violation of

moral norms of handling something or someone. The term "abuse" can also be used in application to the use of substances, but in this paper, our interest is abuse as a type of reprehensible moral action (by the perpetrator) and regrettable/undesirable/experience (by the victim), subject to moral normativity and rationality of emotive reactions (e.g., Strawson, 2008). As such, abuse has a subjective element (e.g., the experience of the treatment; emotional, cognitive, and behavioral reactions to it) and the objective element (e.g., the undue treatment itself, certain types of which, such as gaslighting, are well described in literature).

From the perspective of moral philosophy, abuse is a type of mistreatment that can be characterized in the following way: it involves a violation of moral norms of interpersonal relationships (e.g., between colleagues, friends, romantic partners, relatives, or strangers) and, by extension, a violation of how partners in such relationships are justifiably expected to be treated. Abuse often exploits preexisting power imbalances in such relationships (e.g., seniority at work, parent/child, husband/wife) or creates power imbalances through manipulation (e.g., putting another in a dependent position using intimidation and fear or through creating a false sense of security to catch the victim of guard). Abuse causes non-negligible harm (e.g., Ciurria, 2018; DeHart et al., 2010) to the one being abused, often resulting in a trauma. Abuse tends to have profound detrimental effects on victims' sense of identity and victim's place in the world (e.g., Brison, 2003; Crawford et al., 2009). This justifies, on the one hand, the victim's feeling mistreated and/or violated, and, on the other, the perpetrator being judged morally wrong. There are variations, but abuse tends to be a process rather than an event. This does not mean that it necessarily takes years to unfold (even though it often does); but abuse does tend to involve more than one isolated action.

Abuse occurs in degrees. There are lighter forms (with harms that are relatively easy to amend) and heavier forms (long-lasting effects, persisting psychological damage). Here, we are interested in abuse as an *existential situation*, that is to say, an experience that affects negatively the individual's personality structure (through trauma) and ways in which she is able to relate to herself and the world (the existential aspect) (e.g., Allen, 2011). In this respect, we look at therapy as an existential experience which aims to help the individual to recover from the damage through a constructive (and in this sense, therapeutic) personality change (Warner, 2001). Especially detrimental is psychological abuse which often relies on a set of techniques to alter the victim's self-perception and self-image (cf. Brison, 2003; Miller, 2022) and to condition her to accept and reinforce a certain *narrative of herself*. This narrative aims to undermine her ability to be autonomous. This, among other things, includes: decisions, choices, actions, that is to say, life paths as well as (in)ability to feel pleasure etc (e.g., Ciurria, 2018).

In dealing with abuse, a self-narrative is a key interpretative tool, aimed at understanding trauma and the traumatic event (e.g., Apgar, 1998; Brison, 2003; Harvey et al., 2000; Lindemann, 2001; Marin & Shkreli, 2019; Papathomas & Lavallee, 2012; Saha et al., 2011). However, this interpretative task is more complex than it may appear, since victims of abuse are often confused and disoriented. Work on the self-narrative and trauma-narrative through the therapeutic dialogue helps to clarify, define, and prioritize: who did what, what was done, how did it affect "me", why being affected this way had the effect that it did. Drawing attention to the self-narrative and the therapeutic dialogue that helps to analyze and reconstruct the broken self-narrative, we are putting focus on the questions: What matters for the moral patient/victim? What matters subjectively for the resolution of the trauma? Especially in more complex cases, this is of crucial importance because understanding trauma is only possible in the context of the subjective experiences and reflection of the victim him/herself – there is no external, objective way of knowing a priori how a trauma played out on the symbolic level given the unique subconscious constitution of this particular person.

9. Dialogue, shared subjectivity, and recovery

Especially for the complex cases, co-defining therapy goals to include patients' perspective is of utmost importance. This includes co-defining the script of the therapeutic dialogue since dialogue is a mode in which the cooperation between the patient and the therapist happens. Some commonly defined therapy goals could be rather – so to say – "ego-superficial". We do not mean this in an evaluative sense, but simply to say that sometimes correcting beliefs (e.g., through cognitive therapy) or behavioral patterns could be enough and achieved relatively quickly. However, there are cases when such approaches would simply not be enough or suitable for the patient. Some therapy goals, as felt by the patient, could require intervention into deeper layers of the self. In such cases, co-shaping the dialogue (both its goals and its flow) is especially important. For a patient, the dialogue is a tool of (guided) work with her self-narrative; it is de- and re-construction through reconnecting time slices of his/herself. This often includes connecting to the original traumatic event(s) as they were experience there-and-then (through memory or subjective imagery) and subsequent working by me-here-and-now (under guidance of the therapist) through the negative emotions as me-there-and-then. It also may include quite literary re-writing (or rather re-telling) the invasive self-script (e.g., "you are unworthy", "no one care about you" etc.) and re-conceptualizing the scripts of archetypes to exclude abusive elements (e.g., what "fathering"/"mothering" oneself entails). It also happens in the dialogical context that one gets retrained

in the interpretation of situations, so that they are no longer seen as instance of the original trauma (dealing with "triggers"). Re-orienting expectations from one-self toward one's own goals, which in turn presupposes formulating and spelling out, for oneself, one's own ideal self (e.g., "What matters for me?", "Who am I?") – all these are also very often dialogical activities.

The dialogical relationship between the patient and the therapist is asymmetrical with respect to the autonomy from the traumatic situation and the ability to have an "objective" look at it. By objective in this context, we do not mean "one-fit-all-knowing-what-is-right" point of view, but rather the point of view of functional, non-disrupted humanity. The therapists' role is that of a *guide*, a more autonomous other, who can help find a more objective stance toward (and way out of) the traumatic situation/narrative. Therapists cannot and should not define what you are; neither can or should they do the work for you; but they can provide a reference point of what is "nurturing", "worth of a human being", and morally acceptable (and reprehensible), from outside the disturbed self-narrative. The structural goal is to internalize this reference point and make it a part of new identity/narrative; you no longer need therapy when you can be your own guide. This internalization ultimately is an existential experience and happens in the situation of *intersubjectivity*. How is this intersubjectivity used in the building of narrative? What seems to be essential is the ability of the dialogical other to *relate*, on the existential level, to the experience being in an existential situation and its effects on the *person*. This ability to relate existentially, is based on ontological sameness: there must be at least a theoretical possibility for me to experience it in the same (or in a sufficiently similar) way. This includes the ability to feel the same type of emotions, have the same patterns of thought, feel the same loss – that is to say, the ability of being the same subject of abuse were it to happen to you. This ability then serves as a grounding for the patient's feeling of being understood and validated in her experience, as well as recognized in her subjective experience of trauma.

On another level, this ability to relate but still be outside the patient's traumatic experience itself, allows for the creation of new emotional and cognitive structures through the dialogue that can support the recovery: "Since I relate, I can see what is broken/missing/needs to be in place". The self-narrative is not just a story about yourself; it is the story *you identify with*; it is not something that you narrate but rather a narrative that you live through. The dialogical partner contributes to the reconstruction of the self-narrative through validating certain elements of the patient's narrative that have to do with self-definitions, self-attributes, value rank orders, understanding of personal projects and commitments, and invalidating certain others, for example, when they reflect verbal abuser's scripts.

CAs, being artificial agents, lack this ability to relate, and thus they are unable to provide the rich dialogical environment that is essential to the analysis and rebuilding of the self-narrative, especially in complex cases. Due to their algorithmic (deterministic and inflexible conversing nature) they are unable to adapt to the complexity of the lived experience of a trauma and post-traumatic life. When based on text-generating software, CAs scrap the data-carriers (Babushkina & Votsis, 2022) and generate platitudinal lines in a conversation. As such, they are insufficient to warrant the feelings of being recognized and validated. They also lack the high level cognitive and emotive skills that are required to understand the complexity of one's lived experience of trauma and for reconfiguring one's self-narrative; not in the sense of being "an emotional supporter" that can exhibit a certain emotion emotion or utter a statement but in the sense of being instrumental to the patient's own needs in analyzing, understanding, accepting and re-conceptualizing (especially when taking into consideration the role of free associations in creation of trauma through abuse; this is not universalizable). Given this, conversing with an artificial agent does not provide sufficient therapeutic support. As a result, there is a high risk that it will not lead to the desired therapeutic effect. In some cases, it may even increase risk of retraumatization[19] (e.g., Xygkou et al., 2023), when there is no sufficient support of the objective Other available.

Conclusion

In this paper, we pointed the limits of CAs in therapeutic context and argued that these limits lie in complex cases where the path to recovery is idiosyncratic and cannot be automated. We showed that such cases can be construed as existential situations, where both the fabric of familiarity is shattered and, most importantly, the very sense of the self is confused and requires narrative interventions. Rebuilding oneself and restoring one's place in the (social) world, in such cases, requires careful analysis of the underlying structures of the psychological and experiential make-up; working with symbols, associations, memories; putting names on the events in the past; learning to recognize past and present emotions and deal with them; identifying causes – and this is only a small subset of cognitive operations that require guidance and support from someone outside the trap of the trauma. We also suggested that this type of complex cognitive investment is only possible for another human being, sharing the same existential burden and the same hermeneutical agency as the person undergoing therapy, thus excluding artificial agents as suitable dialogical partners. Our stance in this paper has been from within the philosophy of psychology, reflecting on the normative aspects of therapy and on the concept of the self. We brought in two other normative perspectives, namely: (a) ethics, by focusing on the justified expectations of the patient

from a therapist; and (b) epistemology, by discussing the role that various epistemic processes should play in the therapeutic dialogue (and thus, pointing to the epistemic limitations of CAs). This predominantly normative focus, in turn, limited the extent to which we could engage in the discussion about descriptive aspects of therapeutic practice without getting outside the scope of the paper.

The driving force behind this paper has been a worry that the automation of mental care through AI, aimed at substituting certain cognitive skills that a therapist is required to invest in therapeutic processes, may lead to making the patient, with her health goals and rich idiosyncratic lived experience of trauma, irrelevant to the therapy itself. The need to delineate the limitations of CAs is motivated by the need to preserve the patient's perspective and what matters to her in the therapeutic relationship. Patients dealing with an existential situation are full of open-ended questions, irregularities, and confusion, while often lacking the necessary epistemic tools to interpret, clarify, put the name on etc. Especially in complex cases, therapy takes the form of recreating and rewriting a self-narrative which is a complex epistemic undertaking. It unfolds in a medium of a therapeutic dialogue, guided by the therapist. In this process, therapists play a multiplicity of roles, but with respect to the cognition, the therapist is an epistemic partner, helping to analyze, conceptualize, relate, conclude, and interpret. Delegation and substitution of such epistemic roles cannot be taken lightly, given, on the one hand, the epistemic investment of the therapist as a partner in the cognitive endeavor of the therapy, and, on the other hand, the openness of the one undergoing therapy to the harmful consequences of the inability of artificial agents to fully fulfill those roles.

What our discussion suggests is that CAs in psychotherapy are subject *to epistemo-ethical constraints* (Babushkina & Votsis, 2022): restrictions on their use as a result of their epistemic limitations as "a cognition technology" (Babushkina, 2024). To say that CAs are a cognition technology, is to say that they are designed to act as substitutes of such rich cognitive processes as understanding in the communication with humans and between humans. And as a result, the path to informed and responsible use of such technology lies through thorough critical evaluation of their functionality and ability to fulfill specific goals of human communication in which they are interjected, such as the goals a therapeutic dialogue. Grodniewicz and Hohol (2023) argued that the way CAs are typically marketed suggests that they should be compared to what we expect from a psychotherapist. However, CAs clearly fall short of the comparison because they cannot provide the same rich and flexible therapeutic environment as the one constituted in the relationship between patient and human therapist. This carries a very real risk of "therapeutic misconception". Khawaja and Bélisle-Pipon's (2023) illustrate this risk with a fictional example of Jane: "the main purpose of using the

chatbot was to provide more affordable and accessible therapy to Jane while also assisting her therapist in monitoring her moods so she could provide better care. The therapeutic misconception occurred when Jane misunderstood the limitations of such a technology and overestimated its ability to provide the same therapeutic support and guidance as her therapist would during her in-person sessions. Jane possibly assumed that the chatbot could be utilized as a replacement for traditional therapy. However, that is far from the truth as such chatbots cannot replace human therapists since they lack empathy, curiosity, and connection which are all integral in providing quality care. If users begin to rely on such chatbots as their sole form of therapy, this can have determinantal outcomes such as inadequate support and guidance, which could potentially worsen their mental health".

Whereas there are significant limitations on how CAs can be used in psychotherapy, we do not want to suggest that CAs do not have a place in therapy at all. CAs might be well suited to substitute those elements of therapy that do not rely on intersubjectivity or the ability of the therapist to relate to the existential situation of the patient. For instance, they might – with proper precuations – be relevant for psychoeducation (Grodniewicz & Hohol, 2024), in helping patients to develop certain skills (e.g., personal keeping track of mood changes and thought patterns), in applying therapy insights to specific situations, or in guiding relaxation and breathing exercises.[20] The role of CAs could be similar to those of self-help toolkits (e.g., books, video clips). Studies such as the one conducted by Fitzpatrick et al. (2017) suggest that when treated as such, CAs may even offer some advantage. The authors describe an experiment where a group using a behavioral therapy chat-bot (presented to them as being "closer to a 'choose your own adventure self-help book' and therefore not fully capable of understanding what the needs of the user may be") reported a decrease in symptoms of depression compared to a control group that was provided access to a more traditional e-book on depression. Promising as these results are, it is hard to interpret and extrapolate them. We need more detail, a more systematic understanding of such studies and their results: What are the reasons for observed dynamics? Do the positive effects last? For whom and under what conditions this experience may not have depression-reducing effects? And so, the insertion of CAs into the therapy process should be careful and mindful of the goals it aims to achieve and potential risks for the patient. Special attention should be paid to how CAs are applied: in most cases it might not be harmful or even beneficial to use CAs for psychoeducation, whereas other examples such as the development of so-called deathbots that simulate the deceased and which are developed to cope with grief raise very different concerns. To do justice to the multiplicity of methods and tasks within the complex domain of psychotherapy, further detailed and multi-disciplinary research is

required to determine exactly in what functions LLM-based CAs can safely be employed without undermining therapy goals and violating its norms. We would like, however, to emphasize that one should avoid reliance on silent implicit consent of the patient to the use of CAs with the assumption of their benefit – the patient should be properly informed and have a right to refuse and to withdraw the concern to the use of CAs in her therapy. Patients and therapists need to be adequately informed about the limitations of conversational agents such that they can evaluate the extent to which they indeed contribute to their therapeutic goals.

Notes

1. For an example of technical discussion of LLM in CAs for mental health, see Meyer and Elsweiler (2024).
2. Reviewing the central assumptions of these different approaches is not within the scope of this paper.
3. By "analytical" in this context we mean the standard philosophical method of unwrapping the logical implications of concepts, claims and theories. Our goal is to deduce from the concept and goals of psychotherapy what, rationally speaking, matters for the one in need of psychotherapy. This analysis requires a certain degree of theoretical abstraction, and therefore aims at general principles rather than empirical observations. We will, whenever possible, refer the reader to empirical studies in relevant discussions throughout the paper. However, since LLM-based CAs are a very recent development, there is a lack of in-depth evaluative studies on their success as well as on the patient feedback. We would like to thank the anonymous reviewers for their comments and insights that help us clarify this point as well as the point in the next footnote.
4. We interpret the term "normative" in this context from the standpoint of the philosophical ethics where "justified expectations" refer to rational norms governing therapist-patient relationship. As pointed out by a reviewer, this standpoint is, in a sense, *ideal*-istic, because it aims at grasping the ideal of psychotherapy and its rational norms. Our goal is to understand what the (epistemic and moral) role of therapist *should* be, given justifiable expectations of the one who requires therapy. But that, of course, does not mean to deny that reality is often quite different.
5. See also Grodniewicz and Hohol (2023).
6. This question is, again, asked from a rational standpoint and should be understood not as "Whether mimicking X skills worked for this or that specific case," but rather as "Whether we are justified to believe that mimicking X will be sufficient for everyone's therapy goals and thus can be made a norm/default in therapy"?
7. This is not to deny that there are similarities and common features, but to cast a doubt over one-size-fits-all solution for psychological problems.
8. This is not to say that therapy must be based on verbal dialogue.
9. For a review of narrative approaches in psychotherapy, see Avdi and Georgaca (2007).
10. In the context of the discussion of CAs, it is easy to get an impression that the work on self-narrative should happen in an (all) verbal environment. This is not the case, and we are not making such an assumption here. As a matter of fact, the difficulty or even impossibility to verbalize is often a part of the existential situation. Dialogue, in the way we mean here, does not exclude non-verbal and non-discursive means of

communication between two epistemic agents (patient and therapist), however we will not discuss these in this paper.

11. Not to say that it is necessary for all existential situations.
12. This is not to imply that raising concerns and acting on them (e.g., by changing therapists) is straightforward. Patients are already in a vulnerable position, having to deal with an existential situation. They might lack the confidence to mention their concerns. Furthermore, there might be a lack of alternative therapists available at a given place and time. It is beyond the scope of this paper to detail the negative experiences of patients and the way those affect the success of psychotherapeutic treatment. For an overview of such experiences, see for example Vybíral et al. (2024).
13. It is difficult to draw a clear line between grief and mourning, as grief inevitably involves having a relation with one's socio-cultural environment in which norms and expectations play an important role and inform the grieving process. For the purposes of this paper, we can understand grief and mourning as practically entangled, yet analytically distinct.
14. This is not the place to discuss the key differences between Goldie's and Ratcliffe's accounts. For such a discussion, see: Ratcliffe (2022, pp. 35–42).
15. In this sense, they differ from accounts of grief postulating that grief is causally connected to a distinct object. For Cholbi (2022), for instance, the material object of grief is the "death of a particular individual in whom the bereaved has invested their practical identity" (p. 49), which we grieve for (formal object) because the relationship with the individual is necessarily transformed as are our hopes and expectations with regard how this relationship might have been (2022, p. 57). On this account, without a distinct material and formal object, there would be no grieving process.
16. The DSM-V speaks of *prolonged grief disorder* (PGD) when someone experiences disabling grief for longer than six months to a year. Grieving for such a long period occurs approximately in seven to ten percent of the cases (Szuhany et al., 2021). Although this pathologization of grief is controversial (e.g., Cacciatore & Frances, 2022) and the distinction between typical and pathological grief might not be clear-cut, it seems – especially when it is considered that grief is not the causal result of a particular event – reasonable to say that grief comes in different degrees and in different intensities and that there are differences in how people cope with loss.
17. Moreover, insofar as grief borders on certain mental illnesses – and in fact there are good reasons to assume that at least some symptoms of non-ordinary grief overlap with those of other mental illnesses (Ratcliffe, 2022, p. 193) – it is key for conversational agents to be able to distinguish between issues that are part of the grieving process and those that might have a different origin.
18. For attempts to propose a unified account of psychological abuse, see Follingstad and DeHart (2000); Follingstad (2007).
19. A reviewer has raised an interesting objection that a similar worry can be voiced with respect to a human therapist: an insensitivity and lack of empathy coming from a human may traumatize as much or even more than when it comes from a CA. We think this is a fair point. Furthermore, retraumatization may happen due to factors that are not necessarily connected to therapist's actions; it may occur by chance while coming across elements reminiscent of the traumatic experience. However, it is a part of the therapist's task – in her role *as a therapist*—to create the environment in which such encounters do not reinforce the trauma or lead to a new trauma. In this section we mention retraumatization as a possible consequence of an *epistemic failure* of a therapist (i.e., a failure to help to perform cognitive operations necessary for dealing with trauma, such as analysis, comprehension, drawing links,

identifying causes etc.), not of an *emotive failure* (e.g., to experience an appropriate emotional response to the feelings of others). Our goal is to show CAs' epistemic limitations in therapeutic context and that, while humans have the necessary cognitive abilities, machines do not. There is an increased risk when patients – because of the analogy with human therapists – expect CAs to perform certain epistemic tasks which are beyond their limitations.

20. We have intentionally refrained from mentioning concrete examples of CAs on the market, as we want to avoid involuntary advertising or appearing as promoting specific systems or functionalities. We aimed at remaining on a more general, philosophical level of discussion that leads to normative reflection and a further conversation about guiding principles. With this in view, the paper is limited to suggestions concerning possible uses of LLM-based CAs, given their epistemic limitations. Concrete empirical study, therefore, falls outside our scope. For a reader, who wishes to look deeper into the concrete examples and use cases, it could be interesting to look at more descriptive empirical studies (e.g., J. P. Grodniewicz & Hohol, 2024; Herbener et al., 2024; Khawaja & Bélisle-Pipon, 2023; Wester et al., 2024; Yang et al., 2023).

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Dina Babushkina http://orcid.org/0000-0003-4899-8319

References

Abd-Alrazaq, A., Alajlani, M., Alalwan, A., Bewick, B., Gardner, P., & Househ, M. (2019). An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics*, *132*, 103978. https://doi.org/10.1016/j.ijmedinf.2019.103978

Ahmad, R., Siemon, D., Gnewuch, U., & Robra-Bissantz, S. (2022). Designing personality-adaptive conversational agents for mental health care. *Information Systems Frontiers*, *24*(3), 923–943. https://doi.org/10.1007/s10796-022-10254-9

Aho, K. (2019). *Contexts of suffering: A heideggerian approach to psychopathology*. Rowman & Littlefield International.

Allen, M. (2011). *Narrative therapy for women experiencing domestic violence: Supporting women's transitions from abuse to safety*. Jessica Kingsley Publishers.

Apgar, S. C. (1998). Fighting back on paper and in real life: Sexual abuse narratives and the creation of safe space. In T. Kuribayashi & J. Tharp (Eds.), *Creating safe space: Violence and women's writing* (pp. 47–58). Suny Press.

Avdi, E., & Georgaca, E. (2007). Narrative research in psychotherapy: A critical review. *Psychology & Psychotherapy: Theory, Research & Practice*, *80*(3), 407–419. https://doi.org/10.1348/147608306X158092

Babushkina, D. (2024). *AI, decisions, and the reasons to believe. [manuscript submitted for publication]. Department of behavioural, management and social sciences (philosophy section)*. University of Twente.

Babushkina, D., & Votsis, A. (2022). Epistemo-ethical constraints on AI-human decision making for diagnostic purposes (co-authored with A.Votsis). *Ethics and Information Technology*, *24*(22). https://doi.org/10.1007/s10676-022-09629-y

Bérubé, C., Schachner, T., Keller, R., Fleisch, E., Wangenheim, F., Barata, F., & Kowatsch, T. (2021). Voice-based conversational agents for the prevention and management of chronic and mental health conditions: Systematic literature review. *Journal of Medical Internet Research*, *23*(3), e25933. https://doi.org/10.2196/25933

Brison, S. J. (2003). *Aftermath: Violence and the remaking of a self*. Princeton University Press.

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

Bunz, M., & Braghieri, M. (2022). The AI doctor will see you now: Assessing the framing of AI in news coverage. *AI & Society*, *37*(1), 9–22. https://doi.org/10.1007/s00146-021-01145-9

Cacciatore, J., & Frances, A. (2022). DSM-5-TR turns normal grief into a mental disorder. *The Lancet Psychiatry*, *9*(7), E32. https://doi.org/10.1016/s2215-0366(22)00150-X

Carlbring, P., Hadjistavropoulos, H., Kleiboer, A., & Andersson, G. (2023). A new era in internet interventions: The advent of Chat-GPT and AI-assisted therapist guidance. *Internet Interventions*, *32*, 100621. https://doi.org/10.1016/j.invent.2023.100621

Cholbi, M. (2022). *Grief: A philosophical guide*. Princeton University Press.

Ciurria, M. (2018). The loss of autonomy in abused persons: Psychological, moral, and legal dimensions. *Humanities*, *7*(2), 48. https://doi.org/10.3390/h7020048

Crawford, E., Liebling-Kalifani, H., & Hill, V. (2009). Women's understanding of the effects of domestic abuse: The impact on their identity, sense of self and resilience. *Journal of International Women's Studies*, *11*(2), 5. https://vc.bridgew.edu/jiws/vol11/iss2/5

DeHart, D., Follingstad, D., & Fields, A. (2010). Does context matter in determining psychological abuse? Effects of pattern, harm, relationship, and norms. *Journal of Family Violence*, *25*(5), 461–474. https://doi.org/10.1007/s10896-010-9307-7

Denecke, K., & May, R. (2023). Developing a technical-oriented taxonomy to define archetypes of conversational agents in health care: Literature review and cluster analysis. *Journal of Medical Internet Research*, *25*, e41583. https://doi.org/10.2196/41583

Eshghie, M., & Eshgie, M. (2023). ChatGPT as a therapist assistant: A suitability study. Avaliable at SSRN. https://doi.org/10.2139/ssrn.4423832; https://ssrn.com/abstract=4423832 [Archieved manuscript]

Fiske, A., Henningsen, P., & Buyx, A. (2019). Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *Journal of Medical Internet Research*, *21*(5), e13216. https://doi.org/10.2196/13216

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

Follingstad, D. (2007). Rethinking current approaches to psychological abuse: Conceptual and methodological issues. *Aggression & Violent Behavior*, *12*(4), 439–458. https://doi.org/10.1016/j.avb.2006.07.004

Follingstad, D., & DeHart, D. (2000). Defining psychological abuse of husbands toward wives: Contexts, behaviors, and typologies. *Journal of Interpersonal Violence*, *15*(9), 891–920. https://doi.org/10.1177/088626000015009001

Fuchs, T. (2024). Understanding Sophia? On human interaction with artificial agents. *Phenomenology and the Cognitive Sciences*, *23*(1), 21–42. https://doi.org/10.1007/s11097-022-09848-0

Goldie, P. (2011). Grief: A narrative account. *Ratio*, *24*(2), 119–137. https://doi.org/10.1111/j.1467-9329.2011.00488.x

Graber-Stiehl, I. (2023). Is the world ready for ai-powered therapy? *Nature*, *617*(7959), 22–24. https://doi.org/10.1038/d41586-023-01473-4

Grodniewicz, J., & Hohol, M. (2023). Waiting for a digital therapist: Three challenges on the path to psychotherapy delivered by artificial intelligence. *Frontiers in Psychiatry*, *14*, 1190084. https://doi.org/10.3389/fpsyt.2023.1190084

Grodniewicz, J. P., & Hohol, M. (2024). Therapeutic chatbots as cognitive-affective artifacts. *Topoi*, *43*(3), 1–13. https://doi.org/10.1007/s11245-024-10018-x

Haemmerli, J., Sveikata, L., Nouri, A., May, A., Egervari, K., Freyschlag, C., Lobrinus, J. A., Migliorini, D., Momjian, S., Sanda, N., Schaller, K., Tran, S., Yeung, J., & Bijlenga, P. (2023). ChatGPT in glioma adjuvant therapy decision making: Ready to assume the role of a doctor in the tumour board? *BMJ Health & Care Informatics*, *30*(1), e100775. https://doi.org/10.1136/bmjhci-2023-100775

Harvey, M. R., Mishler, E. G., Koenen, K., & Harney, P. A. (2000). In the aftermath of sexual abuse: Making and remaking meaning in narratives of trauma and recovery. *Narrative Inquiry*, *10*(2), 291–311. https://doi.org/10.1075/ni.10.2.02har

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

Hocking, J., Oster, C., Maeder, A., & Lange, B. (2023). Design, development, and use of conversational agents in rehabilitation for adults with brain-related neurological conditions: A scoping review. *JBI Evidence Synthesis*, *21*(2), 326–372. https://doi.org/10.11124/JBIES-22-00025

Jabir, A., Martinengo, L., Lin, X., Torous, J., Subramaniam, M., & Tudor Car, L. (2023). Evaluating conversational agents for mental health. *Journal of Medical Internet Research*, *25*(1), e44548. https://doi.org/10.2196/44548

Kennedy, A. (2021). *Diagnosis: A guide for medical trainees*. Oxford University Press.

Khawaja, Z., & Bélisle-Pipon, J. C. (2023). Your robot therapist is not your therapist: Understanding the role of AI-powered mental health chatbots. *Frontiers in Digital Health*, *5*, 1278186. https://doi.org/10.3389/fdgth.2023.1278186

Li, H., Zhang, R., Lee, Y.-C., Kraut, R. E., & Mohr, D. C. (2023). Systematic review and meta-analysis of ai-based conversational agents for promoting mental health and well-being. *NPJ Digital Medicine*, *6*(1), 236. https://doi.org/10.1038/s41746-023-00979-5

Lindemann, H. (2001). *Damaged identities, narrative repair*. Cornell University Press.

Marin, K., & Shkreli, A. (2019). An examination of trauma narratives: Narrative rumination, self-reflection, and identity in young adulthood. *Journal of Adolescence*, *76*(1), 139–151. https://doi.org/10.1016/j.adolescence.2019.08.007

McGee, R., & Wolfe, D. (1991). Psychological maltreatment: Toward an operational definition. *Development & Psychopathology*, *3*(1), 3–18. https://doi.org/10.1017/S0954579400005034

Meadows, R., Hine, C., & Suddaby, E. (2020). Conversational agents and the making of mental health recovery. *Digital Health*, *6*, 1–11. https://doi.org/10.1177/2055207620966170

Meyer, S., & Elsweiler, D. (2024). "You tell me": A dataset of GPT-4-Based behaviour change support conversations. In P. Clough, M. Harvey, & F. Hopfgartner (Eds.), *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval* (pp. 411–416). Association for Computing Machinery.

Miller, S. (2022). Toward a relational theory of harm: On the ethical implications of childhood psychological abuse. *Journal of Global Ethics*, *18*(1), 15–31. https://doi.org/10.1080/17449626.2022.2053562

Nicolson, P. (2019). *Domestic violence and psychology: Critical perspectives on intimate partner violence and abuse*. Routledge.

O'Hagan, K. P. (1995). Emotional and psychological abuse: Problems of definition. *Child Abuse and Neglect*, *19*(4), 449–461. https://doi.org/10.1016/0145-2134(95)00006-t

Papathomas, A., & Lavallee, D. (2012). Narrative constructions of anorexia and abuse: An athlete's search for meaning in trauma. *Journal of Loss & Trauma*, *17*(4), 293–318. https://doi.org/10.1080/15325024.2011.616740

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

Prakash, A., & Das, S. (2020). Intelligent conversational agents in mental healthcare services: A thematic analysis of user perceptions. *Pacific Asia Journal of the Association for Information Systems*, *12*(2), 1. https://doi.org/10.17705/1thci.12201

Ratcliffe, M. (2022). *Grief worlds: A study of emotional experience*. The MIT Press.

Saha, S., Chung, M. C., & Thorne, L. (2011). A narrative exploration of the sense of self of women recovering from childhood sexual abuse. *Counselling Psychology Quarterly*, *24*(2), 101–113. https://doi.org/10.1080/09515070.2011.586414

Sarkar, S., Gaur, M., Chen, L., Garg, M., Srivastava, B., Dongaonkar, B. (2023). Drugst.One - a plug-and-play solution for online systems medicine and network-based drug repurposing. *arXiv*. https://doi.org/10.1101/2020.03.22.002386

Scheiderer, E. (2012). Elder abuse: Ethical and related considerations for professionals in psychology. *Ethics & Behavior*, *22*(1), 75–87. https://doi.org/10.1080/10508422.2012.638828

Sedlakova, J., & Trachsel, M. (2023). Conversational artificial intelligence in psychotherapy: A new therapeutic tool or agent. *The American Journal of Bioethics*, *23*(5), 4–13. https://doi.org/10.1080/15265161.2022.2048739

Strawson, P. (2008). *Freedom and resentment and other essays*. Routledge.

Svenaeus, F. (2000). *The hermeneutics of medicine and the phenomenology of health*. Springer.

Szuhany, K., Malgaroli, M., Miron, C., & Simon, N. (2021). Prolonged grief disorder: Course, diagnosis, assessment, and treatment. *Focus*, *19*(2), 161–172. https://doi.org/10.1176/appi.focus.20200052

Tekin, Ş. (2021). Is big data the new stethoscope? Perils of digital phenotyping to address mental illness. *Philosophy & Technology*, *34*(3), 447–461. https://doi.org/10.1007/s13347-020-00395-7

Tolman, R. (1992). Psychological abuse of women. In R. T. Ammerman & M. Hersen (Eds.), *Assessment of family violence: A clinical and legal sourcebook* (pp. 291–310). John Wiley & Sons.

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

Vaidyam, A. N., Wisniewski, H., Halamka, J., Kashavan, M., & Torous, J. (2019). Chatbots and conversational agents in mental health: A review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, *7*(4), 456–464. https://doi.org/10.1177/0706743719828977

Viduani, A., Cosenza, V., Araújo, R. M., & Kieling, C. (2023). Chatbots in the field of mental health: Challenges and opportunities. In I. C. Passos, F. D. Rabelo-da-Ponte, & F. Kapczinski (Eds.), *Digital mental health: A practitioner's guide* (pp. 133–148). Springer.

Vybíral, Z., Ogles, B. M., Řiháček, T., Urbancová, N., & Gocieková, V. (2024). Negative experiences in psychotherapy from clients' perspective: A qualitative meta-analysis. *Psychotherapy Research*, *34*(3), 279–292. https://doi.org/10.1080/10503307.2023.2226813

Warner, S. (2001). Disrupting identity through visible therapy: A feminist post-structuralist approach to working with women who have experienced child sexual abuse. *Feminist review*, *68*(1), 115–139. https://doi.org/10.1080/01417780110042437

Wester, J., Pohl, H., Hosio, S., & van Berkel, N. (2024). "This chatbot would never. . .": Perceived moral agency of mental health chatbots. In J. Nichols (Ed.), *Proceedings of the ACM on human-computer Interaction, 8(CSCW1)* (pp. 1–28). Association for Computing Machinery.

Xygkou, A., Siriaraya, P., Covaci, A., Prigerson, H. G., Neimeyer, R., Siagn Ang, C., & Jou She, W. (2023). The "conversation" about loss: Understanding how chatbot technology was used in supporting people with grief. In A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, A. Peters, S. Muller, J. R. Williamson, & M. L. Wilson (Eds.), *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany* (p. 646). https://doi.org/10.1145/2544548.3581154

Yang, R., Wibowo, S., & Mubarak, S. (2023). An investigation into domestic violence victims' adoption of chatbots for help-seeking: Based on the UTAUT2 and health belief models. *PACIS 2023 Proceedings*, 31. https://aisel.aisnet.org/pacis2023/31