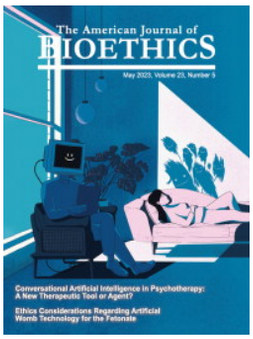

The American Journal of Bioethics

ISSN: 1526-5161 (Print) 1536-0075 (Online) Journal homepage: www.tandfonline.com/journals/uajb20

Conversational Artificial Intelligence in Psychotherapy: A New Therapeutic Tool or Agent?

Jana Sedlakova & Manuel Trachsel

To cite this article: Jana Sedlakova & Manuel Trachsel (2023) Conversational Artificial Intelligence in Psychotherapy: A New Therapeutic Tool or Agent?, The American Journal of Bioethics, 23:5, 4-13, DOI: 10.1080/15265161.2022.2048739

To link to this article: https://doi.org/10.1080/15265161.2022.2048739

Published online: 01 Apr 2022.

THE AMERICAN JOURNAL OF BIOETHICS
2023, VOL. 23, NO. 5, 4–13
https://doi.org/10.1080/15265161.2022.2048739

Taylor & Francis
Taylor & Francis Group

TARGET ARTICLE

Conversational Artificial Intelligence in Psychotherapy: A New Therapeutic Tool or Agent?

Jana Sedlakova[a] and Manuel Trachsel[a,b,c]

[a]University of Zurich; [b]University Hospital Basel; [c]University Psychiatric Clinics Basel

ABSTRACT

Conversational artificial intelligence (CAI) presents many opportunities in the psychotherapeutic landscape—such as therapeutic support for people with mental health problems and without access to care. The adoption of CAI poses many risks that need in-depth ethical scrutiny. The objective of this paper is to complement current research on the ethics of AI for mental health by proposing a holistic, ethical, and epistemic analysis of CAI adoption. First, we focus on the question of whether CAI is rather a tool or an agent. This question serves as a framework for the subsequent ethical analysis of CAI focusing on topics of (self-) knowledge, (self-)understanding, and relationships. Second, we propose further conceptual and ethical analysis regarding human-AI interaction and argue that CAI cannot be considered as an equal partner in a conversation as is the case with a human therapist. Instead, CAI's role in a conversation should be restricted to specific functions.

KEYWORDS

Artificial intelligence; psychotherapy; agency; therapeutic alliance; ethics

INTRODUCTION

According to the World Health Organization (WHO 2019), "[m]ental disorders are one of the most significant public health challenges in the WHO European Region, as they are the leading cause of disability and the third leading cause of overall disease burden." One of the most common mental illness worldwide is depression, whereas about two-thirds of patients are left with unmet needs (WHO 2020, 2017). At the same time, artificial intelligence (AI) and machine learning are rapidly progressing and increasingly applied in mental health care (Burr and Floridi 2020; Torous et al. 2020). This opens previously unimaginable and hardly predictable opportunities and perils.

One of the new technologies is conversational AI (CAI) also known as conversational agents and chatbots—often embodied in mobile applications. Its advertised purpose is to provide mental health support or even solutions very often for depression and anxiety.[1] More precisely, CAI is a software that simulates conversations with users through natural language processing (Adamopoulou and Moussiades 2020). CAI's overarching aims are to help individuals to learn new skills and techniques, implement them in day-to-day life, and recognize behavioral patterns. Thereby, methods of psychotherapeutic treatments such as cognitive-behavioral therapy (CBT), methods from positive psychology, and mindfulness are implemented. Moreover, some conversational agents and chatbots are presented as emotionally intelligent (Mumuksh, Varshney, and Anita 2020; Ghandeharioun et al. 2019) and aim at forming therapeutic alliance with users (Darcy et al. 2021).

The greatest potential of CAI lies in providing care for vulnerable groups such as the elderly, adolescence, and people who do not receive treatment for several reasons (e.g., fear of stigmatization, financial problems, or preference for other solutions than traditional treatments) (Fiske, Henningsen, and Buyx 2020; Luxton 2020). Furthermore, the advances in machine learning and AI can improve the quality of care by patients' empowerment, assisting patients to implement concepts and techniques in the day-to-day life, treatments personalization and identification of mental health problems in early stages due to digital

CONTACT Jana Sedlakova jana.sedlakova@ibme.uzh.ch Institute of Biomedical Ethics and History of Medicine, University of Zurich, Zurich 8006, Switzerland.

[1]The prominent and widely used examples that implement cognitive behavioral therapy are: Flow: https://flowneuroscience.com/home/app/; Woebot: https://woebothealth.com/; Wysa: https://www.wysa.io/

phenotyping (Tekin 2020). The significance of this potential is reflected in the growing research and clinical interest in CAI as well as increasing numbers of new providers and developers (Bendig et al. 2019; Fiske, Henningsen, and Buyx 2020). However, the application of CAI brings also unprecedented challenges and many open questions that need to be inquired to better understand its impact on individuals and society in the long-term perspective.

The present article has two objectives. First, as a framework for the subsequent ethical analysis of CAI, the question whether CAI is rather a tool, or an agent is discussed. The ethical analysis and requirements differ depending on whether it is about a tool or an agent.

Second, we propose further conceptual and ethical analysis regarding human-AI interaction as a framework for defining CAI status and role in the psychotherapeutic setting. More specifically, we conduct the analysis regarding topics of (self-)knowledge, (self-) understanding, and relationships as psychotherapy is embodied in conversation and intersubjectivity. We argue for the thesis that CAI should not be treated as a tool merely implementing evidence-based therapies nor as a digital therapist that can be a partner in a conversation, but its status and role respectively on the spectrum between a tool and a therapist need to be defined. CAI's role in a conversation should be restricted to specifically defined functions and a proper balance of its human-like features should be found.

THE CURRENT STATE OF RESEARCH

Even though millions of people already use "digital therapists," their impact has not been sufficiently evaluated and understood (Shatte, Hutchinson, and Teague 2019). The literature suggests that CAI can offer important benefits and studies have shown promising results regarding its effectiveness for primary prevention, therapy, and relapse prevention mainly for people with mild depression Burr and Floridi 2020; Rubeis 2021; Fiske, Henningsen, and Buyx 2020). Nevertheless, there is precaution needed to make general statements due to the preliminary nature and the early stage of research in this area (Bendig et al. 2019; Thieme, Belgrave, and Doherty 2020; Torous, Cerrato, and Halamka 2019). A related problem is that widely available chatbots and conversational agents are often promoted without empirical evidence (Bendig et al. 2019; Kretzschmar et al. 2019).

Most current scholarly effort in medical ethics of AI-driven mental health applications is focused on issues such as privacy, security, or evidence (Bauer et al. 2020; Luxton 2020; Wang, Fagan, and Yu 2020). Those studies deal with important aspects of ethical issues of new technologies. However, due to the specific focus and immediacy of these issues, such an approach cannot fully address long-term impacts on individuals and society neither the potential of technology to cause major shifts in concepts, behavior, and practices. AI-driven mental health applications are related to a breadth of life aspects and diverse stakeholders. This complexity asks for a more holistic approach (Rubeis 2021) that includes diverse relevant dimensions of AI applications in psychotherapy. In this paper, we will focus on conceptual, epistemic, normative, and ethical dimensions of human-AI interaction that lies at the core of CAI. The inclusion of these dimensions might also strengthen ethical analyses in which the immediate threats are translated into the four principles of biomedical ethics (Beauchamp and Childress 2013). Considering these biomedical principles is salient to place the ethical considerations into the medical context. Nevertheless, the same problem of not addressing far-reaching issues arises. The complexity of the new technology might shift the underlying conception of autonomy, beneficence, non-maleficence, and justice or might introduce new important normative and conceptual distinctions. Causing harm in a traditional setting might mean something different than causing harm in a digital world. For example, the integrity of psychologists is one of the core principles in psychotherapy (APA 2017), and forms of deception are justified only under exceptional circumstances. However, is it justified if a chatbot interacts as if it was empathetic?

In the literature, there are voices calling for more in-depth studies, holistic and human-centred approaches and research focused on long-term societal and individual impacts of the novel technology (Bendig et al. 2019; Burr and Floridi 2020; Rubeis 2021;). Furthermore, researchers have pointed out that there is a lack of ethical guidelines for and criteria of development and application of AI technologies as well as related training of health care professionals. Finally, experts from the field of ethics of AI for mental health are acknowledging the importance of researching how the interaction with CAI can influence the therapeutic relationship, self-understanding, and identity (Burr and Floridi 2020; Fiske, Henningsen, and Buyx 2020). Yet, there is only little existing research in this regard. We want to take these holistic tendencies in the research as a basis from which the long-term effects of CAI can be explored and ethically evaluated.

A CHANGE OF PERSPECTIVE

In what follows, we argue for the following underlying thesis: *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.* This spectrum represents the main theoretical framework in which further analysis focusing on the topics of (self-)knowledge, (self-) understanding and the role of relationship that are central for psychotherapy will be carried out. Furthermore, this theoretical framework reflects two common accounts of AI in ethics and philosophy of technology (Vincent 2020) that have the potential to provide conceptual clarification of the status of CAI and consequently its role in psychotherapy as well as normative requirements.

On the one hand, technology, including AI, can be understood as an *instrument* that helps reach specific ends. On the other hand, some authors advocate for attributing *agency* to AI.[2] For example, James Moor (2006) identifies four stages of AI's moral agency. This distinction between a tool and an agent represents the main theoretical framework for this article because it implies different normative requirements. "A good tool" means something different than "a good agent." To better portray the distinction and its challenges, we will compare one of the tendencies of how CAI is presented in the research with the narratives of its advertisements on the market.[3]

The strong focus of current research on immediate threats caused by new features of technology might imply that CAI is *merely a tool* for familiar practices whose new conditions of use need to be secured (e.g., data protection). In the case of AI-driven chatbots for mental health, this would mean that a chatbot is a mere tool for evidence-based practice of CBT or other approaches of psychotherapy. This is also reflected in the evaluation models and guidelines of mental health apps that focus mostly on the technical side (APA 2021). However, this would ignore the wider implications and impact of CAI.

In the ethics and philosophy of technology, it is uncontroversial that technology and its development is closely linked to values and is re-shaping our environment, understanding, and practices (Burr and Floridi 2020; Stanghellini and Leoni 2020). For example, the use of social media has introduced a new concept of "purported friendship" (Burr and Floridi 2020). CAI can cause major shifts in values, concepts, and generally in the psychotherapeutic landscape (Burr and Floridi 2020; Gabriels and Coeckelbergh 2019). More specifically, the way of how users acquire knowledge and understanding about themselves, the world, and others as well as how they interact and eventually form a reliable relationship with a digital therapist is profoundly different from the interaction with another human. Thus, CAI does not only implement evidence-based therapies, but also has the potential to change them and their effect. Given the transformative character of new technologies, to consider CAI as a mere tool would be illusory and, in the end, it would underrate its potential and impact.

On the contrary, the narratives of how CAI is presented by some developers and providers might give the impression that AI has shifted from being a tool to *being a subject* because of its strong anthropomorphic features. This is mainly the case when chatbots are presented as emotionally intelligent or as having the ability to form a therapeutic bond with users.[4] Opposite to the above account, these narratives seem to overrate the potential and impact of CAI.

It is uncontroversial that CAI does not fulfill the standard requirements to qualify as a subject. In the philosophy of mind, mental states or consciousness, autonomy, and intentionality are often postulated in order to acquire the status of an agent or subject (Schlosser 2019).[5] CAI is a sophisticatedly developed and effective system for data procession and evaluation that shows or mimic some agent-like features, but it would be highly controversial to attribute the above-listed properties to it. Algorithm driven technologies are far away from having mental states and understanding concepts or theories that are often considered to be constitutive for discursive practice and hence a conversation (Green 2020). Thus, it must be reminded that CAI is able *to mimic* a conversation,

[2]In the article, we use the term "agent" as it is used in the philosophical discussion on agency and we will accompany it with the term "subject" to make the reference to human features of agency (e.g., concsiousness and mental states) more explicit. The philosophical concept "agent" also refers to a bearer of moral responsibility and other important moral attributes.

[3]The strategy to take narratives as a starting point is in line with Coeckelbergh's approach (Reijers and Coeckelbergh 2020).

[4]For example, the chatbot Wysa is presented as "AI Chat that makes you feel heard". https://www.wysa.io/

[5]The discussion on subjectivity from philosophy of mind and philosophy of AI is beyond the scope of this paper. In the paper, we focus on the phenomenal level of the interaction with CAI and the ethical consequences. The metaphysical question does not change the argumentation pursued here. The underlying uncontroversial claim is that CAI does not have consciousness and mental states that are necessary for a full attribution of subjectivity and agency (Schlosser 2019).

not to genuinely have it. Finally, to cite famous Searle's response to the Touring test, "[computers/programs] have only a syntax but no semantics" (Searle 1980).

These two approaches toward CAI show that CAI has a hybrid nature. This nature can be summarized in the following three tensions between CAI's tool-like and agent-like features:

1. CAI is a program based on algorithms developed with some purpose that is lacking—among others—mental states and intentionality. Those are tool-like features. At the same time, CAI operates by engaging users in communication. Communication is usually understood as an interaction between two agents. Furthermore, CAI seems to enter our discursive practice.
2. Where is communication there is already a relationship (Kempt 2020). Moreover, therapeutic conversation aims at building a relationship (Miner et al. 2019). This implies some agent-like and social features even though CAI lacks the conditions of being fully attributed as an agent.
3. On the phenomenal level, CAI might be experienced and treated as if it was a subject or agent which, in the end, is the very core of the Touring test and reflects the trends in AI development (Kempt 2020). The more CAI is designed with anthropomorphic traits such as mimicking empathy and emotions, the more it is experienced as another subject even though it is not.

NORMATIVITY AND THE STATUS OF CONVERSATIONAL AI

Depending on where on the continuum between a tool and an agent the CAI is defined and perceived, different normative requirements and expectations can be formulated. The ethical analysis differs depending on whether it is about a tool or an agent. To illustrate this distinction in ethical requirements, we will consider two hypothetical situations. In the first hypothetical situation, CAI has the role of a digital therapist and is perceived as an agent. The second hypothetical situation presents CAI as a tool.

Treating CAI as an *agent* would orientate the ethics toward questions of defining duties, responsibilities or values that are typical for subjects. Furthermore, a formulation of a code of practice parallel to mental health professionals (Fiske, Henningsen, and Buyx 2020) would be a relevant option the more CAI is perceived as an agent. One could go even a step further and claim that a necessary condition for an ethical CAI forming a therapeutic alliance with users would be that values, virtues, and duties parallel to a human therapist are implemented in CAI. To put it differently, if one of CAI's purposes were to establish therapeutic alliance, it would be a reasonable ethical requirement that CAI must fulfill the same ethical conditions as human therapists do. The practical question whether such an implementation is possible will not change soundness of the ethical requirement. However, the practical side whether such an implementation is possible poses an argument against strong inclinations toward CAI as a digital therapist with strong agent-like features. Another consequence of treating CAI as a subject would be that users could expect emotional support and therapeutic alliance basing the human-CAI interaction. In general, the expectations would be much wider and more open than by a tool.

Treating CAI as a *tool* would lead the normative requirements rather into the process of defining specific conditions of safety, reliability, or risk mitigation. Parallel to other tools, specific and limited expectations would make sense (e.g., to get information, mediation of techniques). The interpersonal part would be left for the session with a human therapist.

This strong distinction serves as a method to shed light on the issues that might stay unseen but can have a profound impact. We don't claim that we need to choose between them. Just on the contrary, the hybrid nature of CAI's features asks for finding the right balance between the status of a tool and an agent. If CAI were merely a tool, it would ignore the wider implications because it engages users in conversation, can lead to building a relationship and can be perceived as an agent. If CAI were an agent, it would ignore that CAI is mimicking the conversation, does not have human features like empathy or intentionality and cannot be a bearer of responsibility as humans are. By defining to what extent CAI should or should not be perceived and treated as an agent, normative requirements and thorough guidelines on the development and implementation of CAI can be properly assessed and defined. Such a balance—depending on the needs and context in which CAI is used—aims at developing, integrating, and treating CAI in a responsible way that can benefit patients' well-being at most.

EPISTEMIC PERSPECTIVE AND ITS ETHICAL RELEVANCE: THE AMBIGUITY OF AI

The hybrid nature of CAI poses epistemic and consequently ethical challenges that relate to the ambiguous

status of CAI. We will describe these challenges regarding the topics of (self-)knowledge, (self-)understanding, and relationships.

Conditions of Discursive Practice: (Self-)Knowledge & (Self-)Understanding in a Conversation with AI

One of the important elements of a therapeutic change is the gain of new understanding, knowledge, or insights in the medium of a conversation (Strijbos and Jongepier 2018, Castonguay and Hill 2007). Like a psychotherapeutic process, the interaction with CAI happens in the course of a conversation. However, the nature of such a conversation is profoundly different from a conversation with a human therapist. Therefore, there is a need for a proper ethical and normative analysis to ensure that the interaction with CAI is sufficiently embedded in a normative and ethical framework that aims at patient's beneficence, autonomy, and other health care values. To point out the differences between conversation with a human therapist and CAI, we will analyze the conversation with AI by dint of philosophical theory of pragmatism in which knowledge and understanding are inseparable from agency.

CAI seems to be part of discursive practices because it engages users in conversation, seems to make knowledge claims and it might be perceived as if one had a conversation with it. However, it does not fulfill the requirements of a rational agent that is taking part in discursive practice and hence a conversation. Following Brandom's pragmatic theory, rational agents undertake normative stances such as commitments or entitlements toward their claims (Brandom 2009). That means that a person claiming and believing p is committed and entitled to believe other claims that are inferentially connected with p. On a more general level, one is obliged to integrate new claims and commitments into one's whole and unified belief system. This is possible when rational agents understand and correctly apply concepts. From the practical perspective, one is responsible to offer reasons and can ask for reasons of the claims made by one's counterpart in the conversation. Thus, partners in conversations are engaged in the social and normative game of giving and asking for reasons and are not only rational, but also social and moral agents.

This game is dialectical because conversational partners respect each other, recognize, and acknowledge each other's rational agency and authority of their claims. According to Brandom (2009), being autonomous in a discursive practice means that one can make oneself responsible for one's claims by being able to offer reasons. Furthermore, given the dialectical aspect of discursive practices, one becomes autonomous also by being acknowledged by others as a rational and autonomous agent. Moreover, others can help to gain new knowledge and understanding[6] which might support others in becoming autonomous which plays an important role in psychotherapeutic and medical setting.[7] The importance of intersubjective and social elements for knowledge acquisition is stressed also by other authors. For example, according to Levinas, persons perceive others in their absolute heterogeneity (Gabriels and Coeckelbergh 2019). That means that people are aware that others are out of the scope of their complete understanding. This can have the positive effect that one can be open to new knowledge or point of view while respecting and interacting with others. Another person can give new reasons, widen one's horizon and understanding.

Brandom's conditions of discursive practice and the status of rational agency can be also applied in the context of psychotherapy and could be considered as underlying elements for a psychotherapeutic conversation. The acknowledgement of rational agency and authority of therapists' claims might lead to patients' openness toward novel perspectives that can yield therapeutic change by gathering new knowledge and insights. Therapist's acknowledgment of patient's rational agency and their mutual respect can empower patients' autonomy, agency, and efforts in reaching a therapeutic change.

These conditions of discursive practice don't apply to CAI that cannot be engaged in such a social, dialectic, and normative discursive practice, even though it simulates to do so. One of the main reasons is that CAI does not understand concepts and does not have intentionality. The consequence is twofold. First, CAI cannot explain and offer reasons for its claims in the same way as people. For example, if a user, Peter, asked the CAI why he needs to learn to reformulate his negative thoughts, the CAI could offer different definitions or descriptions of this particular technique. However, the CAI could not truly understand how to explain these in such a way that Peter could better understand his situation, the benefits and purpose of techniques from cognitive behavioral therapy and relate to them. CAI cannot understand and be aware of the situation and reasons why Peter asks and

[6]We use the term "understanding" in its broad sense as a higher epistemic achievement than knowledge (Grimm 2021).

[7]This conception of autonomy fits well the approach of care ethics from biomedical context (Verkerk 2001).

cannot understand what is relevant in such a situation and which individual nuances can make such an explanation successful. CAI works with data that it is fed with and its functions are restricted to mathematical and statistical processes (Coeckelbergh 2020). It would be controversial to claim that CAI can provide robust and complex explanations that might help individual users to better understand their very individual experiences. This problem goes even deeper by considering the black-box problem. The opacity of how algorithms work and how CAI came to a claim or recommendation makes it profoundly problematic to meet the justification or explanation condition. Second, CAI does not have any normative stance and it is implausible to suppose that it could have any without intentionality or understanding. Hence, the entire dialectical process of mutual recognition, respect and acknowledgement that might lead to increased autonomy, new understanding and therapeutic change cannot be present. All in all, rational agency cannot be attributed to CAI. To conclude, the crucial difference between CAI and humans is obvious: CAI mimics being a rational agent, but it is not; therefore, *CAI simulates having a therapeutic conversation, but it does not have any*.

These limitations can have a positive side if the use of CAI is properly defined. CAI might be better at making sound conclusions because it might be less biased and disposed with much more data than a human therapist might be. Thus, CAI's strengths can be in the domains and functions that do not require understanding, empathy, and other human features that cannot be reduced to mathematics or statistics. For example, CAI might be better in recognizing patterns and spotting which interventions and techniques are most successful in particular situations. These functions are different from a psychotherapeutic conversation with a human therapist and must be different because CAI is not a rational and moral agent.

Content of a Conversation with AI: What Kind of (Self-)Knowledge and (Self-) Understanding is Possible?

Not only the conditions for a discursive practice are not fulfilled in a conversation with CAI, the content of such a conversation is also limited and cannot reach the complexity of a therapeutic conversation. The range and kind of data or information provided by CAI might be atypical by comparing it with a knowledge acquisition in a conversation with a human therapist. Firstly, the CAI as an algorithm-driven system is good in providing quantified data or factual information which are limited in range. This type of knowledge can be categorized as third-person knowledge that can inform patients about relationships, human mind, or psychological processes. However, this type of knowledge is insufficient to gain new self-understanding and constitute a therapeutic change. The first-person perspective in which patients can experientially understand their inner states and integrate novel insights in their belief system can constitute a therapeutic change (Strijbos and Jongepier 2018). This type of knowledge and understanding cannot be facilitated by CAI that works only with quantified data and information and lacks important features of a rational, moral, and social agent.

Furthermore, a strong quantification and objectivization of intimate aspects such as emotions, feelings and one's belief system might endanger personal integrity because it might detach people from their qualitative experiences of inner states (Lupton 2016). Instead, the reference to one's inner states can be refocused on numbers and data and not on lived experience and sense-making. In a similar context, Burr and Floridi (2020) have pointed out that such a quantified approach may detach oneself from emotions that are action-guiding by evaluating a situation or environment. Instead, the digital data are only detached records that are available for analysis but are not directly linked to actions or evaluations. Hence, one might tend to identify oneself with numbers and data and in the end, be more detached from the qualitative nature of one's state of mind and other inner states than before. It can have negative consequence for the therapeutic process when patients are trying to reach a therapeutic change and in which the inner states play a prominent role. In general, the salience of the risks is intensified in the psychotherapeutic context in which vulnerable populations use this technology. To conclude, the range of information provided by CAI points to the same distinction: *CAI simulates having a therapeutic conversation, it does not really have it*. To consider CAI's status and role in conversation in a psychotherapeutic setting is crucial because conversation is the essential medium for psychotherapeutic methods and techniques and is embedded in a normative and ethical framework. Thus, it must be ensured that the conversation with CAI is also sufficiently embedded in such a normative and ethical framework to protect patient's beneficence, autonomy, and other important values.

Conclusion: (Self-)Knowledge and (Self-) Understanding in a Conversation with AI

From our analysis, we conclude that CAI must not be treated as an equal partner in a conversation nor as a digital therapist that can facilitate new understanding, insights, and a psychotherapeutic process. In terms of (self-) knowledge acquisition, CAI can provide novel information and data from the third-person perspective. Hence, CAI might be well suited for educational purposes and mediating specific evidence-based techniques and skills. However, the strong human-like features can lead to an illusion that more can happen in the conversation than what is possible. Given the strong human-like features CAI is developed with, it must be prevented that users and patients form wrong expectations such as having a complex conversation and emotional support in which they are understood and can gain new insights. To prevent this, limitations together with expected goals and functions of CAI in a therapeutic context should be made transparent to users. Since information offered by CAI needs further evaluation, the process of sense-making and integration in one's belief system should be left for the sessions with a human therapist. Therefore, CAI should clarify at the beginning of the conversation what is the scope of its interaction and which goals can and cannot be achieved so that users and patients do not expect complex outcomes from the interaction with CAI. Otherwise, patients' and users' autonomy and psychological integrity might be endangered because they access only a restricted scope of knowledge and consequently have opportunity to create only limited understanding and attitudes toward knowledge enabled by CAI. Thus, CAI can have a mediating role between a patient and a human therapist. Its features between being a tool and an agent need to be well balanced as well as communicated to users. This approach acknowledges CAI's broader impact and at the same time, limits its function in psychotherapeutic context to tasks that do not require rational and moral agency.

A Relationship with Conversational AI

The therapeutic relationship is one of the most important elements of a successful psychotherapy (Elsner and Rampton 2020; Torous, Cerrato, and Halamka 2019; Wampold 2015). The interaction with CAI and its human-like features introduces a new concept of digital relationship and with it, new normative questions and challenges regarding a human-AI relationship (Burr and Floridi 2020; Fiske, Henningsen, and Buyx 2020, Bendig et al. 2019). The conditions and norms of a healthy and sustainably benefiting digital relationship are not clear. This problem is intensified with the current efforts to develop CAI that can form therapeutic alliance (Darcy et al. 2021). Such effort pushes CAI development even stronger toward the direction of being an agent and calls for a proper understanding how such a relationship can be truly beneficial to patients and users. As already mentioned, CAI simulates conversation and similarly, it can at most simulate forming or maintaining a therapeutic relationship.

The ambiguity of CAI's status manifests itself also in this area. On the one hand, CAI does not have enough properties of being an appropriate partner in conversation and relationship because it cannot undertake a normative stance and lacks the heterogeneity of humans. On the other hand, CAI has epistemic supremacy in the conversation because it can provide data and analysis of a scale that humans would not be able to.

This poses several ethical problems mainly regarding the beneficence and autonomy of patients and users. First, forming a relationship with CAI can be perceived as deception and can lead to wrong expectations (Kempt 2020). For example, when patients interact with CAI that writes empathetic statements (e.g., "That is really tough to go through this"), positive reinforcement (e.g., "You are so strong to be able to talk about this issue"), and uses other linguistic expressions that strongly mimic therapeutic or human relationship, patients might easily feel in the same way as if they were chatting with a human therapist and hence expect a more profound therapeutic conversation than CAI is able to offer. Due to CAI's limitations of not being a moral and rational agent, CAI cannot offer therapeutic insights and benefits from a profound therapeutic alliance and conversations. It also cannot care for patients. However, if CAI strongly communicates as a human therapist, such wrong expectations can be easily formed even though CAI states that it is only a robot. Another threat might be that users and patients develop dependency and blind trust toward CAI given its 24/7 availability and epistemic supremacy.

Second, the CAI is not a moral subject, hence it is not a subject with duties, responsibilities, or virtues that govern therapeutic relationships. From an ethical point of view, it is problematic to aim at therapeutic alliance if it cannot be ensured that CAI can bear the same responsibilities and duties as human therapists do. For example, if patients form the belief that CAI

cares for them, this belief is obviously false. Even though they perceive that they have a therapeutic relationship with CAI, such a relationship leads to false beliefs and violate values and biomedical principles that shape therapeutic relationships, e.g., fidelity and veracity (Beauchamp and Childress 2013). Moreover, treating someone as a moral subject also means treating people with respect which shapes one's attitude and behavior. The lack of moral agency might lead to different behavior and responses one would have with another human being. The research has already shown that people tend to verbally abuse chatbots and the extent of abuse is also depending on its gender (Brahnam and Angeli 2008). In the context of mental health, one might more easily tend to pretend to do better while interacting with CAI. One could master the usage of CAI like a video game but the application in the everyday life might not happen. Given that CAI is not a moral subject, genuine dialogue and self-revelation might be highly problematic.

Similar to the topic of rational agency, it is highly problematic to present CAI as a companion or with the ability to create relationships since it does not have a responsibility and does not fulfill the requirements of an agent. If users and patients start to believe that CAI cares for them or respect them (Darcy et al. 2021), they form false beliefs which violates values and principles that govern psychotherapy. There is a need for further reflection to define the right balance when CAI shows some human-like features, for example, a sense of humor, to increase engagement and teaching effects. In such a gamification manner, CAI can have a mediating role of important psychotherapeutic tools and techniques. Nevertheless, the interaction should not lead to false beliefs and wrong expectations.

CONCLUSION

We claim that developing and promoting CAI as digital therapists with strong human-like features pose important ethical and epistemic problems. Instead, the right balance of CAI's human-like features should be found. CAI could have a mediating role in which it can engage users and patients in a conversation with restricted and well-defined purposes and goals that are made transparent to users and patients. CAI can mediate well established and evidence-based techniques and information. However, it cannot have an authentic therapeutic conversation because it does not fulfill the requirements of being a rational and moral agent. There is not enough justification for CAI strongly acting as if it was human, for example, by acting as if it was empathetic or forming therapeutic alliance. The vulnerability of people using CAI and intimacy of mental health and psychotherapeutic process call for more precaution.

The risk of negative effects on patients' autonomy and psychological integrity also originates if it is not clear for users that CAI merely mimics conversation and other human-like features. CAI cannot offer authentic facilitation of new self-understanding, perspective, or revelation. CAI can convey information engagingly and entertainingly to decrease retention rate. However, if users' expectation and motivation for using CAI are to have authentic dialogue, gain new self-knowledge or understanding, and feel human closeness, CAI is not suitable for these aims and users might end up in an illusion. The illusion might lead to decreased autonomy. It is not enough when providers explicitly state that CAI is not a human. Knowing that CAI is not human does not automatically mean that users are also aware that CAI is not able to respect them or be part of the authentic dialectical conversation as described above. There is a need for public discussion to increase public awareness and understanding of this new technology.

In general, it is important to further research what effects the novel range and ways of acquiring information and being part of a simulated conversation have on users' and patients' autonomy and psychological integrity. Thereby, it is crucial to gain insights into users' and patients' experiences with CAI. Furthermore, a more constructive approach might be to look at CAI as a *novel type of epistemic exchange* instead of trying to mimic human conversation. The novel type of epistemic exchange asks for new epistemic norms and normative conditions under which patients' well-being can improve most. For this aim, conceptual analysis together with phenomenological insights into patients' experiences with CAI is crucial.

FUNDING

The author(s) reported there is no funding associated with the work featured in this article.

ORCID

Jana Sedlakova http://orcid.org/0000-0002-6887-5941
Manuel Trachsel http://orcid.org/0000-0002-2697-3631

REFERENCES

Adamopoulou, E., and L. Moussiades. 2020. An overview of Chatbot Technology. In *Artificial intelligence applications*

and innovations. AIAI 2020 IFIP Advances in Information and Communication Technology, ed. I. Maglogiannis, L. Iliadis, and E. Pimenidis, 584:373–383. Cham: Springer International Publishing. doi:10.1007/978-3-030-49186-4_31.

American Psychiatric Association (APA). 2021. The App Evaluation Model. https://www.psychiatry.org/psychiatrists/practice/mental-health-apps/the-app-evaluation-model

APA. 2017. Ethical principles of psychologists and code of conduct.

Bauer, M., T. Glenn, J. Geddes, M. Gitlin, P. Grof, L. V. Kessing, S. Monteith, M. Faurholt-Jepsen, E. Severus, and P. C. Whybrow. 2020. Smartphones in mental health: A critical review of background issues, current status and future concerns. *International Journal of Bipolar Disorders* 8 (1):2. doi:10.1186/s40345-019-0164-x.

Beauchamp, T. L., and J. F. Childress. 2013. *Principles of biomedical ethics*. 7th ed. Oxford: Oxford University Press.

Bendig, E., B. Erb, L. Schulze-Thuesing, and H. Baumeister. 2019. The next generation: chatbots in clinical psychology and psychotherapy to foster mental health – A scoping review. *Verhaltenstherapie* 2019:1–13. doi:10.1159/000501812.

Brahnam, S., and A. D. Angeli. 2008. Special issue on the abuse and misuse of social agents. *Interacting with Computers* 20 (3):287–91. doi:10.1016/j.intcom.2008.02.001.

Brandom, R. 2009. *Reason in philosophy: Animating ideas*. Brandom: Harvard University Press.

Burr, C., and L. Floridi. 2020. *Ethics of digital well-being*, vol. 140, Cham: Springer International Publishing.

Castonguay, L. G., and C. E. Hill. 2007. *Insight in psychotherapy*. Washington, DC: American Psychological Association.

Coeckelbergh, M. 2020. *AI Ethics*. Cambridge, MA: The MIT Press.

Darcy, A., J. Daniels, D. Salinger, P. Wicks, and A. Robinson. 2021. Evidence of human-level bonds established with a digital conversational agent: Cross-sectional, retrospective observational study. *JMIR Formative Research* 5 (5):e27868. doi:10.2196/27868.

Elsner, A. M., and V. Rampton. 2020. Ethics of care approaches in psychotherapy. In *Oxford handbook of psychotherapy ethics*, ed. M. Trachsel, J. Gaab, N. Biller-Andorno, Ş. Tekin, and J. Z. Sadler. Oxford: Oxford University Press.

Fiske, A., P. Henningsen, and A. Buyx. 2020. The implications of embodied artificial intelligence in mental healthcare for digital wellbeing. In *Ethics of digital well-being*, ed. Ch. Burr and L. Floridi, 207–219. Cham: Springer International Publishing.

Gabriels, K., and M. Coeckelbergh. 2019. Technologies of the self and other: How self-tracking technologies also shape the other. *Journal of Information, Communication and Ethics in Society* 17 (2):119–27. doi:10.1108/JICES-12-2018-0094.

Ghandeharioun, A., D. McDuff, M. Czerwinski, and K. Rowan. 2019. EMMA: An emotion-aware Wellbeing Chatbot. *International Conference on Affective Computing*. http://arxiv.org/pdf/1812.11423v2.

Green, M. 2020. Speech Acts. *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/win2020/entries/speech-acts/

Grimm, S. 2021. Understanding. *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/sum2021/entries/understanding/

Kempt, H. 2020. *Chatbots and the Domestication of AI*. Cham: Springer International Publishing.

Kretzschmar, K., H. Tyroll, G. Pavarini, A. Manzini, and I. Singh. 2019. Can your phone be your therapist? Young people's ethical perspectives on the use of fully automated conversational agents (Chatbots) in Mental Health Support. *Biomedical Informatics Insights* 11: 1178222619829083. doi:10.1177/1178222619829083.

Lupton, D. 2016. *The quantified self*. Cambridge: Polity Press.

Luxton, D. 2020. Ethical implications of conversational agents in global public health. *Bulletin of the World Health Organization* 98 (4):285–7. doi:10.2471/BLT.19.237636.

Miner, A. S., N. Shah, K. D. Bullock, B. A. Arnow, J. Bailenson, and J. Hancock. 2019. Key considerations for incorporating conversational ai in psychotherapy. *Frontiers in Psychiatry* 10:746. doi:10.3389/fpsyt.2019.00746.

Moor, J. H., 2006. The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems* 21(4): 18–21. doi:10.1109/MIS.2006.80

Mumuksh, B., R. Varshney, and R. Anita. 2020. Emotionally intelligent ChatBot for mental healthcare and suicide prevention. *International Journal of Advanced Science and Technology* 29 (6):605–2597.

Reijers, W., and M. Coeckelbergh. 2020. *Narrative and Technology Ethics*. Cham: Springer International Publishing.

Rubeis, G. 2021. E-mental health applications for depression: An evidence-based ethical analysis. *European Archives of Psychiatry and Clinical Neuroscience* 271 (3): 549–55. doi:10.1007/s00406-019-01093-y.

Schlosser, M. 2019. Agency. *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/win2019/entries/agency/

Searle, J. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3 (3):417–57. doi:10.1017/S0140525X00005756.

Shatte, A. B. R., D. M. Hutchinson, and S. J. Teague. 2019. Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine* 49 (9): 1426–48. doi:10.1017/S0033291719000151.

Stanghellini, G., and F. Leoni. 2020. Digital phenotyping: Ethical issues, opportunities, and threats. *Frontiers in Psychiatry* 11:473. doi:10.3389/fpsyt.2020.00473.

Strijbos, D., and F. Jongepier. 2018. Self-knowledge in psychotherapy: Adopting a dual perspective on one's own mental states. *Philosophy, Psychiatry, & Psychology* 25 (1): 45–58. Project MUSE, doi:10.1353/ppp.2018.0008.

Tekin, Ş. 2020. Is big data the new stethoscope? Perils of digital phenotyping to address mental illness. *Philosophy & Technology* 34 (3):447–461. doi:10.1007/s13347-020-00395-7.

Thieme, A., D. Belgrave, and G. Doherty. 2020. Machine learning in mental health. *ACM Transactions on Computer-Human Interaction* 27 (5):1–53. doi:10.1145/3398069.

Torous, J., K. J. Myrick, N. Rauseo-Ricupero, and J. Firth. 2020. Digital mental health and COVID-19: Using technology today to accelerate the curve on access and quality tomorrow. *JMIR Mental Health* 7 (3):e18848. doi:10.2196/18848.

Torous, J., P. Cerrato, and J. Halamka. 2019. Targeting depressive symptoms with technology. *mHealth* 5:19. doi: 10.21037/mhealth.2019.06.04.

Verkerk, M. A. 2001. The care perspective and autonomy. *Medicine, Health Care and Philosophy* 4 (3):289–94. doi: 10.1023/A:1012048907443.

Vincent, M. C. 2020. Ethics of artificial intelligence and robotics. *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/win2020/entries/ethics-ai/

Wah, B. W. 2009. *Wiley encyclopedia of computer science and engineering*. 5th ed. Hoboken, NJ: John Wiley.

Wampold, B. E. 2015. *The great psychotherapy debate: The evidence for what makes psychotherapy work*. 2nd ed. New York, NY: Routledge.

Wang, L., C. Fagan, and C. Yu. 2020. Popular mental health apps (MH apps) as a complement to telepsychotherapy: Guidelines for consideration. *Journal of Psychotherapy Integration* 30 (2):265–73. doi:10.1037/int0000204.

WHO. 2017. *3 out of 4 people suffering from major depression do not receive adequate treatment*. https://www.euro.who.int/en/media-centre/sections/press-releases/2017/3-out-of-4-people-suffering-from-major-depression-do-not-receive-adequate-treatment

WHO. 2019. *Mental Health: Fact Sheet*. https://www.euro.who.int/__data/assets/pdf_file/0004/404851/MNH_FactSheet_ENG.pdf

WHO. 2020. *Depression: Key Facts*. https://www.who.int/news-room/fact-sheets/detail/depression