Conversational AI for Psychotherapy and Its Role in the Space of Reason

JANA SEDLAKOVA
Institute for Biomedical Ethics and History of Medicine, University of Zurich

1. INTRODUCTION

The developments of artificial intelligence (AI) re-open and re-frame many traditional philosophical questions such as what rationality, reasoning or free will is or what it means to be a human being. With the recent development of ChatGPT technology, these questions pertained to the public space and steered a discussion particularly about the reasoning capabilities of conversational and generative AI in comparison to humans and its role in our rational discourse. These discussions are closely linked to the topic of the possibility of artificial generative intelligence (GAI). The recent book by Landgrebe and Smith (2022) offers compelling arguments against the possibility of GAI as well as machines' abilities to master human language, social interaction and morality. Despite these arguments, there is a problem on the side of human's imaginative power to perceive more than there is and treat AI as humans and social actors (Banks 2019; Nass and Moon 2000) independent of its actual properties and abilities or lack of those. The mathematical and ontological arguments will not help against this strong human tendency to treat conversational AI (CAI) as if it was human. This tendency is reinforced by the fact that CAI is developed with the aim to appear human-like. The consequence of this might be that on the phenomenological level and pragmatically speaking, AI could be acknowledged to master language, enter our discursive practice and be a social actor despite its lack of human properties. Perhaps, it is something that we are already seeing with the current version of ChatGPT. I will argue that this phenomena needs to be taken seriously.

I will focus on a specific domain of CAI's application, namely, healthcare because the importance of the gap between CAI's actual properties and emulated ones; and consequently ethical considerations are particularly important given the highly sensitive context and vulnerable groups that use CAI. This is intensified when CAI is used for psychotherapeutic purposes. Hence, the understanding of its capabilities and limitations as an agent entering the psychotherapeutic and rational discourse is essential. In the context of healthcare, CAI can be used for such purposes as real-time data collection, answering patients' questions, providing them with information and accompanying or initial interventions (Ahmed et al. 2021; Fiske et al. 2019). The wide use and importance of CAI are reflected in the fact that this technology can be certified as a medical device. For example, the chatbots for mental health Woebot and Wysa were recognized by FDA as medical devices. The increased recognition of this technology reflects its poten-

tially important role in healthcare, mainly considering closing the treatment gap and providing an underdeserved population with initial interventions.

In this paper, I will offer a reflection on CAI and a way of making sense of it. Thereby, I will place the reflection in a broader context of pragmatism and the narratives accompanying it inspired by postphenomenological account of technology (Ihde 1990; Verbeek 2005). Both place their subject—either language or technology—in a broader context of social and normative practices. In this understanding, technology is a mediator between humans and the world and is influenced by those practices, including the language used when describing it (Coeckelbergh 2018, 2020a). Thus, one of the important aspects of understanding technology is looking at the narratives and framings encompassing it (Coeckelbergh 2018; Mager and Katzenbach 2020)—this will be the starting point. Based on this, I will analyze CAI's role and limits in the space of reasons, mainly by drawing on Brandom's philosophy accompanied by reflections inspired by Wittgenstein and Searle. Finally, I will draw conclusions regarding CAI's limits and role in psychotherapy and mental health.

2. THE HUMANIZATION NARRATIVE OF CAI

CAI in the context of mental health and well-being is often described as "digital therapists". Already this description gives a sense of one of the main narratives regarding this technology. CAI is developed with human-like features and with the aim to emulate human features and abilities such as intelligence, reasoning or empathy. The humanization narrative and idea date back to Alan Turing and his famous imitation game or also called the Turing test (Turing 1950). Despite the strong trends of developing human-like AI, it is worth asking if the concept of Turing test is meaningful in the context of CAI. Every narrative has a guiding power because it is a framework which allows some options and does not give way to others. It is a framework that nurtures some questions and omits others. The humanization narrative leads to understanding AI in human terms instead of trying to make sense of AI per se with its own peculiarities. In the end, AI is a new entity with new capabilities and limitations. If the humanization narrative is too strong it can have a negative effect of a Procrustean bed. This can be illustrated by the recent strong trend in the development of CAI in mental health and well-being—namely to develop it in a way that CAI is able to form a therapeutic relationship with its users (Darcy et al. 2021). The therapeutic alliance is even measured by the same instruments as with human therapists. Is this useful, valuable and desirable? The research suggests that the human-likeness might be an important factor in positive experiences and adherence (Abd-Alrazaq et al. 2019). However, the studies are highly inconsistent in their measurements and theoretical underpinnings (Li and Suh 2021). This is not surprising because CAI is often measured as if it was a human therapist and there is not enough conceptual understanding that would adequately capture its role and place in our practices and its actual properties that should be measured.

The consequence of these humanization trends leads to giving users and patients the narrative of human-likeness and hence at least implicitly guiding them in using and treating this technology parallel to practices with a human therapist or physician. Consequently, users are encouraged to form a therapeutic relationship with it and understand their experiences with CAI similar to an experience with a human. Due to the interactive nature, users might tend to over-trust this technology and the pieces of advice that it might give to them (Chow et al. 2023). This is problematic because CAI's responses depend on the formulation of the problem or questions (Chow et al. 2023; Dowling 2023). The problem is intensified when CAI gives wrong and harmful health advice. For example, the National Eating Disorders Association needed to take their chatbot, Tessa, offline because it provided users with harmful responses (McCarthy 2023). Hence, the humanization aspects of the interaction with CAI are connected with many ethical problems and the narrative opens questions such as: should we trust digital therapists the same way as we trust human therapists? If we change the perspective by going outside of this narrative, then we could ask:

Does it even make sense to create human-like AI? What is the actual value of creating human-like AI and emulating human properties and abilities?

Perhaps this narrative of human-like AI arises from an overoptimistic and non-reflective understanding of AI that is, in the end, not (always) helpful. In this context, it might be important to understand philosophical work as therapeutic work in line with Wittgenstein (Wittgenstein 1953)—shedding light on our misconceptions and ways how we create meaning in our practices. What is needed is to find novel ways of conceptualizing CAI that will be more informative not only for using the CAI and making sense of it in an individual experience but also for study designs and measurements. In the rest of the paper, I would like to analyze CAI in terms of its possible role in the space of reasons and consequently its limits there. Thereby, I do not attempt to provide a full account of CAI's limitations and strengths, just to sketch a path in the complex landscape of CAI and its application in psychotherapy and mental health care.

3. CAI IN THE SPACE OF REASON

When talking to CAI, it seems like CAI can offer information, explanations even empathetic statements. In the case of CAI for psychotherapy and mental health, CAI teaches evidence-based techniques, often based on cognitive behaviour therapy and can offer insides into someone's life. For example, CAI can spot such patterns in behaviour as a tendency to have lower moods on specific days. Users can ask for advice and help. What cannot be denied is that we talk to CAI—we engage with it in a conversation, in discursive practice. In the case of digital therapists, users even engage with it in a therapeutic conversation and a therapeutic relationship (Darcy et al. 2021). At least, it seems like it. However, in which sense is CAI part of this practice?

Pragmatic epistemological theories understand language and discursive practice as embedded in a social and normative space (Brandom 2009; McDowell 1984). Being part of discursive practice means having certain commitments and entitlements (Brandom, 2009). Brandom formulates this in the following way:

> [Knowing] involves adopting three different attitudes: attributing a commitment, attributing an entitlement, and undertaking a commitment. […] Knowledge is intelligible as a standing in the space of reasons, because and insofar as it is intelligible as a status one can be taken to achieve in the game of giving and asking for reasons. But it is essentially a social status, because it incorporates and depends on the social difference of perspective between attributing a commitment (to another) and undertaking a commitment (oneself) (Brandom 1995, pp. 903-904).

When making a claim, we are committed to offering reasons and are consequently vulnerable to criticisms because others attribute the same commitment to us. When making a claim, we, at least implicitly, attribute entitlement to others. Namely, we expect that others will endorse the claim and use it in the space of reasons as well. In return, we are acknowledged by others as reliable and competent rational agents with some level of epistemic authority and trustworthiness. The aspect of otherness could be added to the space of reasons. We talk to others that can have a different perspective based on differences in understanding, attitudes and experiences (Gabriels and Coeckelbergh 2019; Strijbos and Jongepier 2018; Zahavi 2014).

Similar to the speech act theory (Searle 1969), an important aspect of this account is that participating in a space of reason is not about empirical description of mental states such as knowing, but about actively participating in a normative practice and maintaining it (Heinrichs and Knell 2021) by undertaking commitments. To formulate it with Wittgensteinian terms, it is about making a move in the game of giving and asking for reasons and knowing which moves are allowed and which are not. This game is social.

Coming back to the initial question: is CAI making a move in this social and normative game?

It is and it is not. It seems like CAI makes claims, maybe even knowledge claims (Heinrichs and Knell 2021). When a digital therapist writes: "This sounds like a busy and stressful time. You decided to show up and take care of your needs. You are showing commitment to your mental health, I am proud of you", it seems like it is expressing utterances about the state of the world and the user. The consequence might be that the user feels entitled to claims made by CAI (Heinrichs and Knell 2021) and form a relationship with it because of the expressive speech acts. Furthermore, it seems like CAI entertains its epistemic authority

(Chow et al. 2023). However, CAI is not part of the space of reason because it is not undertaking and attributing commitments and entitlements. It is not social and normative. It also does not perform expressive speech acts. The moves are not based on semantics, on understanding or normative attitudes, but on statistics, on probabilities. CAI might be able to follow rules, e.g., when the AI is combined with statistical and symbolic rule-following models (Maruyama 2021). However, the decisive question here is the old question of normativity. What does it mean to be following a rule and what type of normativity is constitutive in the space of reason?

I do not attempt to provide an answer to this highly complex question. Instead, I attempt to spark reflections about these traditional questions and express doubts that reasoning and normativity can be reduced to a set of explicitly defined rules (Coeckelbergh 2020b). Furthermore, I argue that it would be simplistic to claim that CAI is not entering the space of reason and is merely a tool because it does not function like us humans and does not have human properties. There are two reasons. First, I argue that in order to make sense of CAI, there is a need for a shift from strongly comparing it to humans because of the danger of a Procrustean bed approach. Instead, having a more creative approach by looking at its peculiarities, strengths and limits can be more helpful in making sense of it. Second, even if CAI does not have the necessary properties of entering the space of reasons, it emulates that it does. The power of this emulation should not be underestimated. More powerful than CAI's actual capabilities is our imaginative power triggered by the emulation of CAI's human-likeness and its role in the space of reason. It is a common and well-known feature of humans to anthropomorphize technology (Nass and Moon 2000; Silvio 2010). However, this tendency is so actively and strongly encouraged by human-like design of CAI that it might be hard to see the difference. If we strongly believe that CAI is like us, it might become like us because we will treat it like it.

The arguments formulated by Landgrebe and Smith (2022) do not seem to account for this phenomenon when CAI is acknowledged as an agent in our discursive practices. They formulate four criteria to be fulfilled for AGI to master a language. The first reads as follows:

> [T]he machine has the capability to engage in a convincing manner with one or more human interlocutors in conversations of arbitrary length in such a way that the human interlocutors do not feel constrained in the realisation of their conversation-related intentions by the machine-interlocutor. This means that when the human interlocutor engages in the conversation, she must be able to realise her intentions without making the sorts of special effort which (for the moment at least) we are familiar with making when dealing with a machine […] (Landgrebe and Smith 2022, pp. 217-218).

It could be argued that human interlocutors will get used to the type of conversation with AI that first was perceived as limiting and different from other human interlocutors, however, with time, it became natural. The feelings of constraint and effort might change with time as they pragmatically depend on the current practices. We can get used to talking to CAI and adapt our conversation style to it. Even though it does not change the fact that CAI does not have human capabilities, it will change the fact that CAI is part of the discursive practice. If only looking at the ontology, many difficulties that are connected to this will be left unaddressed.

To conclude, CAI's emulation of human conversation brings the focus back to us and our practice of how we treat CAI and how we want to treat it. It might be reasonable to shift from the type of questions "Can CAI be part of the space of reasons?" to "Under which conditions is it desirable and beneficial that we treat CAI as if it was part of the space of reasons and in which sense?" Consequently, there is a large set of questions that we need to answer in order to define CAI's role in our practices. These questions include:

What kind of epistemic authority should we ascribe to CAI?
How much or what types of agency should and do we want to ascribe to CAI?
What is the value of a emulation of human-likeness and rational normative discourse?

What type of relationships are desirable to have with CAI?

In the last part, I want to reflect on CAI and its role as digital therapists from an approach inspired by medicine. Medical knowledge has been highly enriched by learnings from pathology. Similarly, looking at pathologies—at mistakes and blind spots—of CAI can inform us about its functionalities and strengths that might be important for our rational and normative discourse as well as therapeutic landscape.

4. THE PATHOLOGIES, STRENGTHS AND LIMITS OF CAI

Earlier this year, Kellin Pelrine beat AI in the Go game with a strategy that would be easily spotted by humans.[1] This shows an important blind spot or pathology of AI, namely its limited capacity for generalization, understanding novel situations and ascribing meaning to them. Floridi and Chiriatti (Floridi and Chiriatti 2020) have described three important tests in which GPT-3 failed in similar ways. Here again, the problem is the absence of AI's understanding because AI works based on statistical models and its output is "a statistically good fit" (Floridi and Chiriatti 2020). In the same article, Floridi and Chiriatti highlight that the issue of AI is not about the output, but about the process of how this output is achieved. This is crucial because the focus on the output might give more power to the emulation of CAI' as a partner in the space of reasons and might turn the emulation into an illusion.

Another pathology of CAI is its lack of otherness, of a unique perspective that each human has (Walsh 2016). The claims or statements that CAI presents are based on a collection of digital data that are statistically processed. What kind of perspective is it? It might seem that it could be the view from nowhere, however, in a negative sense. The nowhere is empty, and cannot bring the novelty and heterogeneity that others do. These are also essential elements of empathy—the ability to access others' experiences and recognize them in their otherness (Irarrázaval and Kalawski 2022). In psychotherapy, empathy plays an important role in helping clients to develop greater empathy towards themselves and others (Irarrázaval and Kalawski, 2022). When chatbot's responses are perceived as empathetic (or even more empathetic than the ones expressed by humans) (Montemayor et al. 2022) the important sphere of intersubjectivity is not present. The chatbot is fully there for the person, however, can it also teach the client to transfer this kind of empathy from "nowhere" to themselves and others? Furthermore, as stated in the previous subchapter, CAI cannot attribute and undertake commitments and entitlements which are essential normative limitations. Finally, an important pathology of CAI is its incapability to distinguish between truth and fiction—this problem is often referred to as the hallucination (Alkaissi and McFarlane 2023) of CAI which once again leads to the humanization narrative.

These pathologies point out important limitations of CAI, but also its important strengths. The underlying statistical processes give CAI the ability to process a large amount of data and spot patterns that humans cannot. This can be useful in the therapeutic process where CAI can offer users insights based on a big variety of data, answer many basic questions about mental health and provide them almost immediately with overviews of current approaches and theories (if properly trained and updated). The limits are its lack to offer reasons and justification for its claims. This calls for caution when ascribing CAI epistemic authority and asking for advice. When integrating CAI's claims into one's own belief system, one solution could be to reflect upon its justification and invite other people, human therapists, to participate in this justification.

The further limits are that CAI is not able to understand the uniqueness of personal experiences and provides users with more profane advice than what can be statistically found in the data. The lack of otherness positively means that CAI cannot be judgmental and it can have the positive effect that users might feel more comfortable sharing their experience (Vaidyam et al. 2019). This is already an important step in a psychotherapeutic process similar to leading a journal. However, the lack of otherness vastly limits CAI's capabilities as a therapist because it cannot offer a second-person perspective that is filled with attitudes such as compassion, empathy or kindness which are important for a therapeutic change (Strijbos and Jongepier 2018). Finally, CAI's lack of normative attitudes makes it unfit as a partner in a relationship, particularly in a

therapeutic relationship that is strongly embedded in a space of values. The power of emulation can be that users can practice techniques and skills in an effective medium where interpersonal aspects are emulated. However, it has be reminded that a emulated relationship and therapeutic process are different from an authentic one.

CONCLUSION

CAI's role in the space of reasons, in our practices such as therapeutic process should be carefully shaped by as many stakeholders as possible and by looking at CAI's own limitations and strengths. The human-likeness of CAI and its human narratives can be dangerous because they might give too much power to the emulation of human-likeness. The danger here is that the emulation might dictate reality and not vice versa. What we need is a more granulate understanding of such concepts as agency, rationality and normativity. We need to refine our understanding of our practices because there is a new entity entering them. Or rather, we allow a new entity to enter them. In the case of psychotherapy, CAI's role should be defined by its strengths and limited by its weaknesses. A careful approach is particularly essential in such a sensitive context as the one of mental health and well-being. There are no digital therapists, there are only emulations of digital therapists. But we can discover the important strengths of these emulations.

NOTES

1 First reported by *The Financial Times*: https://www.ft.com/content/175e5314-a7f7-4741-a786-273219f433a1?uuid=SNeychiOfriWpmcP0682

REFERENCES

Abd-Alrazaq, A. A., Alajlani, M., Alalwan, A. A., Bewick, B. M., Gardner, P., and Househ, M. 2019. An overview of the features of chatbots in mental health: A scoping review. *Int. J. Med. Inf.* 132, 103978. https://doi.org/10.1016/j.ijmedinf.2019.103978

Ahmed, A., Ali, N., Aziz, S., Abd-Alrazaq, A .A., Hassan, A., Khalifa, M., Elhusein, B., Ahmed, M., Ahmed, M. A. S., and Househ, M. 2021. A review of mobile chatbot apps for anxiety and depression and their self-care features. *Comput. Methods Programs Biomed.* Update 1, 100012. https://doi.org/10.1016/j.cmpbup.2021.100012

Alkaissi, H. and McFarlane, S. I. 2023. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus* 15, e35179. https://doi.org/10.7759/cureus.35179

Banks, J. 2019. A perceived moral agency scale: Development and validation of a metric for humans and social machines. *Comput. Hum. Behav.* 90:363-371. https://doi.org/10.1016/j.chb.2018.08.028

Brandom, R., 1995. Knowledge and the Social Articulation of the Space of Reasons. *Philos. Phenomenol. Res.* 55:895-908. https://doi.org/10.2307/2108339

_____. 2009. *Reason in Philosophy: Animating Ideas*. Cambridge, MA: Belknap.

Chow, J. C. L., Sanders, L. and Li, K. 2023. Impact of ChatGPT on medical chatbots as a disruptive technology. *Front. Artif. Intell.* 6, 1166014. https://doi.org/10.3389/frai.2023.1166014

Coeckelbergh, M. 2018. Technology Games: Using Wittgenstein for Understanding and Evaluating Technology. *Sci. Eng. Ethics* 24:1503-1519. https://doi.org/10.1007/s11948-017-9953-8

_____. 2020a. When Machines Talk: A Brief Analysis of Some Relations between Technology and Language. https://doi.org/10.48417/TECHNOLANG.2020.01.05

_____. 2020b. *Ai Ethics*. Cambridge, MA: MIT Press.

Darcy, A., Daniels, J., Salinger, D., Wicks, P. and Robinson, A. 2021. Evidence of Human-Level Bonds Established With a Digital Conversational Agent: Cross-sectional, Retrospective Observational Study. *JMIR Form. Res.* 5, e27868. https://doi.org/10.2196/27868

Dowling, C. 2023. Defining the Future of Open AI in Clinical Care. *Harv. Med. Sch*. https://hcp.hms.harvard.edu/news/defining-future-open-ai-clinical-care.

Fiske, A., Henningsen, P. and Buyx, A. 2019. Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy. *J. Med. Internet Res*. 21, e13216. https://doi.org/10.2196/13216

Floridi, L. and Chiriatti, M. 2020. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds Mach*. 30:681-694. https://doi.org/10.1007/s11023-020-09548-1

Gabriels, K. and Coeckelbergh, M. 2019. 'Technologies of the self and other': how self-tracking technologies also shape the other. *J. Inf. Commun. Ethics Soc*. 17:119-127. https://doi.org/10.1108/JICES-12-2018-0094

Heinrichs, B. and Knell, S. 2021. Aliens in the Space of Reasons? On the Interaction Between Humans and Artificial Intelligent Agents. *Philos. Technol*. 34:1569-1580. https://doi.org/10.1007/s13347-021-00475-2

Ihde, D. 1990. Technology and the Lifeworld: From Garden to Earth. https://philarchive.org/rec/IHDTAT-3.

Irarrázaval, L. and Kalawski, J. P. 2022. Phenomenological considerations on empathy and emotions in psychotherapy. *Front. Psychol*. 13.

Landgrebe, J. and Smith, B. 2022. *Why machines will never rule the world: Artificial intelligence without fear*. London and New York: Routledge.

Li, M. and Suh, A. 2021. Machinelike or Humanlike? A Literature Review of Anthropomorphism in AI-Enabled Technology. In: *Proceedings of the Annual Hawaii International Conference on System Sciences*. Hawaii International Conference on System Sciences. https://doi.org/10.24251/hicss.2021.493

Mager, A. and Katzenbach, C. 2020. Future imaginaries in the making and governing of digital technology: Multiple, Contested, Commodified. https://doi.org/10.31235/osf.io/zhjwk

Maruyama, Y. 2021. Symbolic and Statistical Theories of Cognition: Towards Integrated Artificial Intelligence. In: Cleophas, L. and Massink, M. (eds.), Software Engineering and Formal Methods. SEFM 2020 Collocated Workshops, *Lecture Notes in Computer Science*, pp. 129-146. Cham: Springer. https://doi.org/10.1007/978-3-030-67220-1_11

McCarthy, L. 2023. A Wellness Chatbot Is Offline After Its 'Harmful' Focus on Weight Loss. *New York Times*.

McDowell, J. 1984. Wittgenstein on following a Rule. *Synthese* 58:325-363.

Montemayor, C., Halpern, J. and Fairweather, A. 2022. In principle obstacles for empathic AI: why we can't replace human empathy in healthcare. *AI Soc*. 37:1353-1359. https://doi.org/10.1007/s00146-021-01230-z

Nass, C. and Moon, Y. 2000. Machines and Mindlessness: Social Responses to Computers. *J. Soc*. 56:81-103. https://doi.org/10.1111/0022-4537.00153

Searle, J. R. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

Silvio, T. 2010. Animation: The New Performance? *J. Linguist. Anthropol*. 20:422-438.

Strijbos, D. and Jongepier, F. 2018. Self-Knowledge in Psychotherapy: Adopting a Dual Perspective on One's Own Mental States. *Philos. Psychiatry Psychol*. 25:45-58. https://doi.org/10.1353/ppp.2018.0008

Turing, A. M. 1950. Computing Machinery and Intelligence. *Mind* 59:433-60. https://doi.org/10.1093/mind/lix.236.433

Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S. and Torous, J. B. 2019. Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. *Can. J. Psychiatry Rev. Can. Psychiatr*. 64:456-464. https://doi.org/10.1177/0706743719828977

Verbeek, P.-P. 2005. *What Things Do: Philosophical Reflections on Technology, Agency, and Design*. University Park: Pennsylvania State University Press.

Walsh, P. 2016. Dan Zahavi: Self and Other: Exploring Subjectivity, Empathy, and Shame. *Husserl Stud*. 32. https://doi.org/10.1007/s10743-015-9180-6

Wittgenstein, L. 1953. *Philosophical investigations*. Oxford: Basil Blackwell.

Zahavi, D. 2014. *Self and Other: Exploring Subjectivity, Empathy, and Shame*. Oxford: Oxford University Press.