10.12740/APP/170132

Five reasons why a conversational artificial intelligence cannot be treated as a moral agent in psychotherapy

Marcin Paweł Ferdynus

Abstract

Sedlakova and Trachsel present an analysis regarding the evaluation of a new therapeutic technology, namely conversational artificial intelligence (CAI) in psychotherapy. They suggest that CAI cannot be treated as an equal partner in the therapeutic conversation, because it is not a moral agent. I agree that CAI is not a moral agent. However, I believe that CAI lacks at least five basic attributes or abilities (phenomenal consciousness, intentionality, ethical reflection, prudence, conscience) that would allow it to be defined as a moral agent. It seems that the ethical assessment of the possibilities, limitations, benefits and risks associated with the use of CAI in psychotherapy requires a determination of what CAI is in its moral nature. In this paper, I attempt to show that CAI is devoid of essential moral elements and hence cannot be treated as a moral agent.

conversational artificial intelligence; moral agent; psychotherapy

The development of artificial intelligence (AI) technologies presents new issues for psychiatric ethics [1-6]. For instance, Sedlakova and Trachsel present an analysis regarding the evaluation of a new therapeutic technology, namely conversational artificial intelligence (CAI) in psychotherapy [7]. They suggest that CAI cannot be treated as an equal partner in the therapeutic conversation, because it is not a moral agent. I agree that CAI is not a moral agent (like human). However, I believe that CAI lacks at least five basic attributes or abilities (phenomenal consciousness, intentionality, ethical reflection, prudence, conscience) that would allow it to be defined as a moral agent. Sedlakova and Trachsel do not devote much attention to explicating these properties [7]. It seems that the ethical assessment of the possibilities, limitations, benefits and risks associated with the use of CAI requires a determination of what CAI is in its essence (i.e., its moral nature). In this paper, I attempt to show that CAI is devoid of essential moral elements and hence cannot be treated as a moral agent. I maintain that only a moral agent can provide adequate therapeutic assistance to the patient. I believe that certain moral qualities and abilities are the basis for creating decent conditions for helping other people. Although CAI cannot be recognised as a moral agent, it can be a valuable tool to support the therapeutic process, but it should be used under the supervision of a human therapist. What kind of moral qualities does CAI lack? I start with an example.

When I examine ivory, I not only see its shape but also directly perceive its colour: a mix of yellow, white and cream. While the weight, density and chemical composition of ivory can be examined and described, the situation is different with colour. People cannot perceive my experience of the colour yellow-white-cream. This

Marcin Paweł Ferdynus: The John Paul II Catholic University of Lublin; Lublin, Poland
Correspondence address: marcin.ferdynus@wp.pl

example demonstrates that in human cognition, there are subjective elements that humans know only for themselves because access to those elements is reserved only for those individuals. Even if my brain was emulated (scanned) while I was viewing ivory, no one but me would see what I see and how I see it. These qualities or features are referred to as phenomenal states (*qualia*), and they include, for example, experiencing the blue sky, experiencing mourning, experiencing remorse, or feeling guilty. In addition to the experienced qualities being subjective, they are also unique and unrepeatable. This is evidenced by the fact that it is difficult to describe a colour to a blind person or the experience of a strong feeling or moral state to someone who has never experienced it. True intelligence presupposes phenomenal consciousness (consciousness from a first-person perspective), while CAI lacks it [8]. Thus, CAI is devoid of an essential attribute possessed by moral agents (first reason).

Another problem is that CAI lacks intentionality. Intentionality is a feature of mental states that gives them content, refers to something, concerns something, or is directed at something beyond them. We usually think that for cognition to occur, the existence of beings capable of perceiving and objects perceived by them is sufficient. Franz Brentano noticed that something else appears in our cognition, namely the relation directed at the object. The foundation of this relationship is a feature of mental states (i.e., intentionality), thanks to which our cognition concerns something or is about something. The philosopher John Searle studied this problem in the context of AI. Based on a thought experiment (called Chinese room), he concluded that the formal calculations performed by AI alone are unable to produce intentionality [9]. Searle most likely meant that the meanings attributed to AI programs come solely from human users or programmers. It can be said that the meanings are arbitrary to the program itself, which is semantically empty [8]. If CAI is devoid of understanding and if true intelligence presupposes understanding, then CAI does not possess the essential property that is attributed to moral agents (second reason).

Another difficulty is that CAI is unable to make the ethical reflection that would allow for a morally right choice. Computer scientists who create AI expect clear guidelines from ethicists on how to proceed in a particular case. In other words, programmers demand unambiguous answers from ethicists; that is, the answers leave no moral doubts. Such answers cannot always be given – an unambiguous solution to moral dilemmas does not exist (e.g., trolley dilemma) [10]. There are different reasons for this situation. First, each answer formulated based on a given ethical theory (e.g., consequentialism or deontology) is supported by strong arguments, and the rules of conduct developed on their basis may lead to different decisions. Second, the very choice of a specific ethical theory means that an attempt to solve a moral problem based on this theory implies several possible solutions [11]. Moreover, the choice of one or another ethical theory is already a moral choice, an acknowledgement of beliefs about what is morally good/right and bad/wrong. Third, our relationship to the situation leads us to issue different responses when we respond as uninvolved observers as opposed to active participants. Finally, in some situations, making a morally right decision requires replacing one moral principle with another moral principle. Strict adherence to one ethical principle at all times can lead to the acceptance of an inherently harmful pattern of moral conduct: puritanism [12]. Notably, decision-making in the moral sphere consists of a multi-faceted approach to the situation and an independent determination of the rules that should apply in a particular case. Only someone who is a moral agent is capable of such acts – someone who is not only able to make moral decisions and formulate and understand moral judgements but also able to take responsibility for their own actions. In this sense, CAI is not a being capable of making moral decisions. Therefore, CAI is devoid of another essential property that moral agents possess (third reason).

CAI is also devoid of an important moral component related to the activities of practical reason: it lacks practical wisdom (prudence). Aristotle emphasises that prudence is a permanent character trait/disposition or something that can demonstrate what is best and most perfect in practice [13]. Prudence is an expression of the spiritual maturity of the acting subject, the knowledge of people and world affairs, the ability to search for optimal behaviour in complicated life situations within the limits of applicable moral norms, the flexibility to make a justifiable compromise and firmness in defending the impassable limits of morality. In the virtue of prudence, the vital wisdom of humans manifests itself with special force. AI, founded on self-learning programs, can, for example, bluff in a game of poker, but this is an adaptive action from choosing an effective strategy based on the analysis of opponents' moves. On the basis of such an analysis, it is not possible to make a morally right choice of means in a specific situation. In the case of AI, only a decision based on the efficiency category is possible [14]. If CAI lacks prudence (practical wisdom) and therefore some kind of moral intelligence, then a reasonable assumption is that it lacks another essential quality that moral agents possess (fourth reason).

Conscience plays an important role in the human decision-making process. In the light of the general assessment or norm, conscience is a formed judgement about the moral goodness or badness of a person's own specific act, the implementation of which becomes, for that person, a source of internal approval or a sense of guilt, being a good or bad person. Since conscience determines the moral value of a particular action, indicating at the same time the obligation to perform or omit the action, the ethical function of conscience is normative. This means that conscience formulates specific norms of conduct. Its specificity is expressed primarily in the fact that it is a product of the specific subject in relation to their own specific act. It is recognised that conscience is a subjective and concrete norm, always present in an individual form, enclosed within the moral self-awareness of the acting subject. In short, conscience is the tribunal before which humans are responsible for their own acts. CAI is not endowed with moral self-awareness (conscience) and therefore does not have the ability to formulate practical judgements about moral goodness and badness. Furthermore, CAI is unable to take responsibility for its actions. Thus, CAI is deprived of another important property that characterises moral agents (fifth reason).

I do not claim that the above-mentioned attributes exhaust the set of elements that characterise moral subjects. I argue, however, that they constitute certain minimum conditions that should be met for CAI to be considered a moral agent. The literature suggests that the solution to this troublesome situation could be to equip AI with an 'artificial conscience' and 'artificial practical wisdom' [15]. By equipping CAI with 'artificial moral properties', one could create an 'artificial moral agent'. Even if such a being could be constructed, behind artefacts such as 'artificial conscience' and 'artificial practical wisdom', there will still be algorithms that are a better or worse imitation of human moral attributes. Even assuming that a person could be 'therapeutically happy' with this new technology, their dignity would be deceived in a seemingly harmless way. The deception in this case is that CAI cannot mean what it says, nor can it have feelings for the person. Some people may want or even like to communicate with CAI about existential matters, but the conversation would not be real. In addition, it seems that in emotionally difficult situations, a person wants recognition for their courage, suffering, loss, and harm, not a superficial simulation of compassion [8,16-17]. The clear lack of relevant moral qualities that humans possess does not allow CAI to be recognised as a moral agent. Nevertheless, CAI is a valuable tool supporting the therapeutic process, but it should be used under the supervision of a human therapist.

Acknowledgements

I would like to thank Anna Golova for her valuable comments when preparing this paper for publication. This article was written in part during my stay at the Oxford Uehiro Center for Practical Ethics.

Funding source

This work was supported by the National Science Centre, Poland, under grant no. 2022/06/X/HS1/00470.

REFERENCES

1. Miner AS, Shah N, Bullock KD, Arnow BA, Bailenson J., Hancock J. Key considerations for incorporating conversational AI in psychotherapy. Frontiers in Psychiatry. 2019; 10:746.
2. Hatherley JJ. Limits of trust in medical AI. J Med Ethics. 2020;46(7):478-481.

3. Rubeis G. E-mental health applications for depression: An evidence-based ethical analysis. European Archives of Psychiatry and Clinical Neuroscience. 2021;271(3):549–55.
4. Kempt H, Nagel SK. Responsibility, second opinions and peer-disagreement: ethical and epistemological challenges of using AI in clinical diagnostic contexts. J Med Ethics. 2022;48(4):222-229.
5. Darcy A, Daniels J, Salinger D, Wicks P, Robinson A. Evidence of human-level bonds established with a digital conversational agent: Cross-sectional, retrospective observational study. JMIR Formative Research. 2021;5(5):e27868.
6. Luxton D. Ethical implications of conversational agents in global public health. Bulletin of the World Health Organization. 2020;98(4):285-287.
7. Sedlakova J, Trachsel M. Conversational Artificial Intelligence in Psychotherapy: A New Therapeutic Tool or Agent? Am J Bioeth. 2022;1-10.
8. Boden M. Artificial Intelligence: A Very Short Introduction. Oxford: Oxford University Press; 2018.
9. Searle J. Minds, Brains and Programs. Behav Brain Sci. 1980;3:417-457.
10. Cathcart T. The Trolley Problem or Would You Throw the Fat Guy off the Bridge? A Philosophical Conundrum. Workman Publishing; 2013.
11. Bostrom N. Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press; 2016.
12. Szulczewski G. Artificial intelligence and moral intelligence. An introduction to cybernetic ethics. Ethics in Economic Life. 2019;22:19-31.
13. Aristotle. The Great Ethics of Aristotle. Trans. by P. Simpson. New York: Routledge; 2014.
14. Russell SJ, Norvig P. Artificial Intelligence: A Modern Approach. 3rd edition. Upper Saddle River. NJ: Persons Education Limited; 2016.
15. Lekka-Kowalik A. Morality in the AI World. Law and Business. 2021;1(1):44-49.
16. Ferdynus MP. Albert Mieczysław Krąpiec's theory of the person for professional nursing practice. Nurs Philos. 2020;21(2):e12286.
17. Ferdynus MP. Four philosophical images of man and nursing from Krąpiec's perspective. Nurs Philos. 2021;22(2):e12344.