Conversational Artificial Intelligence and the Potential for Epistemic Injustice

De Proost, Michiel; Pozzi, Giorgia

DOI
10.1080/15265161.2023.2191020

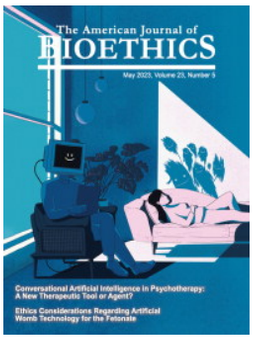
Publication date
2023

Document Version
Final published version

Published in
The American journal of bioethics : AJOB

Citation (APA)
De Proost, M., & Pozzi, G. (2023). Conversational Artificial Intelligence and the Potential for Epistemic Injustice. *The American journal of bioethics : AJOB*, *23*(5), 51-53. https://doi.org/10.1080/15265161.2023.2191020

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.



Taylor & Francis
Taylor & Francis Group

The American Journal of Bioethics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uajb20

Conversational Artificial Intelligence and the Potential for Epistemic Injustice

Michiel De Proost & Giorgia Pozzi

To cite this article: Michiel De Proost & Giorgia Pozzi (2023) Conversational Artificial Intelligence and the Potential for Epistemic Injustice, The American Journal of Bioethics, 23:5, 51-53, DOI: 10.1080/15265161.2023.2191020

To link to this article: https://doi.org/10.1080/15265161.2023.2191020

THE AMERICAN JOURNAL OF BIOETHICS
2023, VOL. 23, NO. 5, 51–53
https://doi.org/10.1080/15265161.2023.2191020

Taylor & Francis
Taylor & Francis Group

OPEN PEER COMMENTARIES

Conversational Artificial Intelligence and the Potential for Epistemic Injustice

Michiel De Proost[a] and Giorgia Pozzi[b]

[a]Ghent University; [b]Delft University of Technology

In their article, Sedlakova and Trachsel (2023) propose a holistic, ethical, and epistemic analysis of conversational artificial intelligence (CAI) in psychotherapeutic settings. They mainly describe interesting challenges regarding the ambiguous therapeutic relationship with its introduction and recommend "conceptual analysis together with phenomenological insights into patients' experiences" (Sedlakova and Trachsel 2023, 11). To contribute to this important reflection, our considerations aim to show that in order to provide a holistic framework of ethical and epistemological issues brought about by CAI in psychotherapy, an analysis of their potential to cause epistemic injustice cannot be left out of the picture. We map out issues in terms of testimonial and hermeneutical injustice that can emerge in connection with CAI and that deserve further attention.

The concept of epistemic injustice was introduced by the feminist philosopher Miranda Fricker (2007) in her landmark monograph *Epistemic Injustice: Power and the Ethics of Knowing*. With this book, she highlighted specific forms of injustice at the intersection of ethics and epistemology. According to Fricker's definition, epistemic injustice is a "wrong done (to) someone specifically in their capacity as a knower" (2007, 1). She identifies two such patterns of wrongdoing as "testimonial injustice" and "hermeneutical injustice." Testimonial injustice takes place at the communication level when certain prejudices and stereotypes cause a hearer to assign a deflated level of credibility to a speaker's testimony. Hermeneutical injustice precedes communication and occurs when a gap in collective interpretative resources puts a person or group at a disadvantage when trying to make sense of their social experiences. This is either due to a lack of concepts to express one's experience (e.g., experiencing postpartum depression before this very concept was collectively available) or to a misalignment between a person's experience and the existing socially accepted concepts (Mason 2021; Pozzi 2023a).

The original Frickerian framework has been refined and elaborated in the context of medicine and healthcare (Kidd and Carel 2017) and also of psychiatric practice (Kidd et al. 2022; Sanati and Kyratsous 2015), although its role in medical AI has received less attention as many discussions on epistemic injustice only focus on human conversations (Pozzi 2023b). We contend that the algorithms used for CAI in psychotherapy could create the perfect storm for epistemic injustice. Because of that, there are good reasons to pause before CAI tools are adopted too widely or permanently in psychotherapy.

Sedlakova and Trachsel (2023) argue that CAI easily can get "epistemic supremacy in the conversation because it can provide data and analysis of a scale that humans would not be able to" (10). It is not hard to imagine how testimonial injustice can occur when persons seek help for their mental health by engaging in communication with CAI. If we prioritize CAI over human dialogue, persons may gradually lose confidence in themselves as epistemic agents because they can get the feeling of not being heard, especially when the chatbot fails to pick up on the user's exact meaning. The limit of the scope of algorithms and their epistemic opacity can create a mismatch between the tracked data or the app's prediction of their experience and their actual experience (Symons and Alvarado 2022). One cannot understand what goes on, even though one knows there is something concealed in an overall hybrid structure that is difficult to espy. Users can thus experience an unjust deflation of their credibility owing to their perceived inferiority to an automated data analytics system. This can be the case, particularly if a human professional entering the

CONTACT Giorgia Pozzi g.pozzi@tudelft.nl Faculty of Technology, Policy and Management, Delft University of Technology, The Netherlands.

therapeutic relationship at a later stage gives more credibility to the assessment provided by the CAI rather than to the patient's testimony. This scenario is not too far-fetched considering the potential of automation bias raised by AI systems introduced in crucial decision-making scenarios (Pozzi 2023b). Against this background, patients may even lose the motivation to talk and share their experiences. This is particularly likely to occur if a patient needs to share sensitive information regarding their mental condition, such as thoughts of self-harm or suicidal ideation. Therefore, a situation that Dotson (2011) described as "testimonial smothering" can emerge in which an epistemic subject, tired of being continually placed at a disadvantage, decides to remain silent and hide their experience altogether. This could potentially end in even more suffering.

Relatedly, there are also wider hermeneutical implications, which we want to draw attention to as well. Sedlakova and Trachsel (2023) correctly recognize that "CAI as an algorithm-driven system is good in providing quantified data or factual information *which are limited in range*" [our emphasis, 9]. Using CAI systems in therapeutical settings, patients' experience needs to fit the (limited) conceptual categories already encoded in the algorithm to be successfully engaged with. As such, a patient's experience that *exceeds* these categories cannot receive appropriate consideration by the CAI involved in the therapeutic conversation. That is to say, if a concept cannot capture a particular patient's experience because it is not part of the system's "vocabulary" in the first place, it will necessarily remain unacknowledged. The consequences are potentially extremely harmful to patients. They will not be able to make sense of their lived experience, and the probability of suffering a hermeneutical marginalization is considerable. In fact, the patient's experience becomes unintelligible to the human operator who will enter the therapeutic relationship and to the patient herself.

This danger of hermeneutical marginalization is even more significant for people belonging to disadvantaged and underrepresented social groups. It is widely acknowledged that AI systems tend to mirror what is best represented in the training data. So, it does not seem unrealistic to say that the lived experiences of underrepresented social groups are likely to be neglected. These considerations pave the way to hermeneutical injustice and point to an unwarranted hermeneutical privilege taken up by the CAI (Pozzi 2023a). The situation briefly depicted is particularly problematic in the context of mental health, in which having the feeling of being acknowledged and receiving an explanation of what one is experiencing are important epistemic tools to understand and hopefully overcome a problematic situation that can be burdensome to the patient.

Let us note that the epistemic injustices we refer to emerge even if CAI systems are included in the therapeutic relationship only in its initial stages as a tool to have a first interaction with patients before they are redirected to a human professional. In fact, deciding which information is *relevant* for diagnostic purposes is already an extremely value-loaded process that needs to take into consideration the contextual aspects that characterize every patient in their singularity. If what a particular patient is experiencing cannot be recognized by the CAI as a "relevant" symptom for a certain mental health condition (because, for example, the person cannot properly articulate their experience due to its ineffability), then the risk of downplaying and not detecting issues that would need attention in a therapeutical setting is considerable.

The present response has focused on what is perceived as a big shortcoming in the article of Sedlakova and Trachsel: the potential of CAI to induce epistemic injustice. We are convinced that the original Frickerian framework and related ameliorative work can be pivotal in unveiling subtle forms of injustice that are potentially going unnoticed in the current CAI debate in the field of psychotherapy. In their conclusion, Sedlakova and Trachsel (2023) point out that the introduction of CAI in therapeutic settings is to be understood in terms of a "novel type of epistemic exchange" (11). However, in the face of the issues pointed out in this commentary, to what extent can we talk about a proper exchange with this kind of technology? Since CAI has limited, pre-determined options on how to react to patients' input, further research is needed to assess whether a patient can genuinely participate in the interaction with CAI in an epistemically relevant sense. Our considerations of possible issues in terms of testimonial and hermeneutical injustice raise significant doubts in this respect and pave the way to further research, especially if we take into account that the vast majority of these tools have not been subjected to empirical scrutiny. It is crucial to keep in mind these limitations, and we see (feminist) bioethicists as uniquely positioned to cut through the hype surrounding CAI and clear the blurred epistemo-ethical dimensions that ought to be considered.

FUNDING

Michiel de Proost's contribution to this work was supported by the H2020 European Research Council (grant number 949841). Giorgia Pozzi's contribution to this work was supported by the European Commission through the H2020-INFRAIA-2018-2020/H2020-INFRAIA-2019-1 European project "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" (Grant Agreement 871042). The funders had no role in developing the research and writing the manuscript.

ORCID

Michiel De Proost http://orcid.org/0000-0003-0545-8515
Giorgia Pozzi http://orcid.org/0000-0001-8928-5513

REFERENCES

Dotson, K. 2011. Tracking epistemic violence, tracking practices of silencing. *Hypatia* 26 (2):236–57. doi:10.1111/j.1527-2001.2011.01177.x.

Fricker, M. 2007. *Epistemic injustice: Power and the ethics of knowing*. Oxford: Oxford University Press.

Kidd, I. J., and H. Carel. 2017. Epistemic injustice and illness. *Journal of Applied Philosophy* 34 (2):172–90. doi:10.1111/japp.12172.

Kidd, I. J., L. Spencer, and H. Carel. 2022. Epistemic injustice in psychiatric research and practice. *Philosophical Psychology* 1–29. doi:10.1080/09515089.2022.2156333.

Mason, R. 2021. Hermeneutical injustice. In *The Routledge handbook of social and political philosophy of language*, eds. J. Khoo and R. K. Sterken. New York: Routledge.

Pozzi, G. 2023a. Automated opioid risk scores: A case for machine learning-induced epistemic injustice in healthcare. *Ethics and Information Technology* 25 (1):3. doi:10.1007/s10676-023-09676-z.

Pozzi, G. 2023b. Testimonial injustice in medical machine learning. *Journal of Medical Ethics*. Published Online First: 12 January 2023. doi:10.1136/jme-2022-108630.

Sanati, A., and M. Kyratsous. 2015. Epistemic injustice in assessment of delusions. *Journal of Evaluation in Clinical Practice* 21 (3):479–85. doi:10.1111/jep.12347.

Sedlakova, J., and M. 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.

Symons, J., and R. Alvarado. 2022. Epistemic injustice and data science technologies. *Synthese* 200 (2):87. doi:10.1007/s11229-022-03631-z.

THE AMERICAN JOURNAL OF BIOETHICS
2023, VOL. 23, NO. 5, 53–55
https://doi.org/10.1080/15265161.2023.2191041

Taylor & Francis
Taylor & Francis Group

OPEN PEER COMMENTARIES

Check for updates

Responsible Use of CAI: An Evolving Field

Mehrdad Rahsepar Meadi, Neeltje Batelaan, Anton J. L. M. van Balkom, and Suzanne Metselaar

Amsterdam UMC location Vrije Universiteit Amsterdam

Sedlakova and Trachsel (2023) argue that on the one hand, conversational artificial intelligence (CAI) does not fulfill the necessary conditions for full attribution of agency, such as having consciousness, mental states and intentionality. On the other hand, they argue, CAI should neither be seen as merely a tool, as this would ignore the wider implications and transformative impact of CAI that are a result of its agent-like features. These features include engaging in communication with the user, building a relationship, and anthropomorphic traits such as mimicking empathy and emotions. Mimicking agency could lead to morally undesirable effects, related to shortcomings in facilitating (self-)understanding and in maintaining a therapeutic relationship as compared to the human therapist. The authors argue that by *simulating* a conversation and by *simulating* having a therapeutic relationship CAI gives "(…) an illusion that more can happen in the conversation than is possible." This could lead the patient to form false beliefs and wrong expectations, which violates the values and principles of psychotherapy.

The authors offer two ways to mitigate the morally undesirable effects of CAI mimicking agency. First,

CONTACT Mehrdad Rahsepar Meadi m.rahseparmeadi@ggzingeest.nl Department of Psychiatry, Amsterdam Public Health, Mental Health program, Amsterdam UMC location Vrije Universiteit Amsterdam, Boelelaan 1117, Amsterdam, The Netherlands.