salud mental

EDITORIAL
Volume 46, Issue 6, November-December 2023
doi: 10.17711/SM.0185-3325.2023.036

Reflections on the use of artificial intelligence in psychiatric diagnosis

Marcos F. Rosetti[1]

[1] Instituto Nacional de Psiquiatría Ramón de la Fuente Muñiz, Ciudad de México, México.

Correspondence:
Marcos F. Rosetti
Instituto Nacional de Psiquiatría Ramón de la Fuente Muñiz, Ciudad de México, México.
Calz. México-Xochimilco 101,
Col. San Lorenzo Huipulco,
Tlalpan, 14370 CDMX, México.

Citation:
Rosetti, M. F. (2023). Reflections on the use of artificial intelligence in psychiatric diagnosis. *Salud Mental*, *46*(6), 285-286.

DOI: 10.17711/SM.0185-3325.2023.036

I recently had the opportunity to read the article "Evaluating the Performance of ChatGPT in Differential Diagnosis of Neurodevelopmental Disorders: A Pediatricians-Machine Comparison," published in *Psychiatry Research* journal. Authors Wei, Cui, Wei, Cheng, & Xu (2023) used Cohen's Kappa values to compare the diagnostic accuracy of a sample of pediatric residents, experienced pediatricians, and ChatGPT-4 (Open AI, San Francisco, California), an artificial intelligence chatbot. In this exercise, the researchers gave the study groups the results of the Early Language Milestone Scale, the Gesell Developmental Scale, the Modified Checklist for Autism in Toddlers, and the Autism Behavior Checklist, as well as the gender and age of patients. During the second round, this information was complemented by vignettes containing aspects of clinical interest such as chief complaint, developmental milestones achieved and family history. For each scenario, both pediatricians and the chatbot were asked to select the most likely diagnosis from Autism Spectrum Disorder, Global Developmental Delay, and Developmental Language Disorder. The study reported diagnostic accuracy values of 66.7% for experienced pediatricians and ChatGPT, surpassing pediatric residents, who achieved just 55.3%. The level of agreement between experienced pediatricians and ChatGPT, according to the Kappa value, was .43. When vignettes were included, the chatbot's accuracy decreased to 53.3%, and interobserver agreement between the chatbot and pediatricians dropped to .35. Interestingly, this exercise was performed with the ChatGPT-4 version available to the public (at the time of writing, it operated under a subscription model) and is only trained with non-specialized information available on the Internet. The next few years will see many more articles like this one, particularly since it is possible to specialize the linguistic models under which these types of tools operate using data sets with better resolution or specificity to answer questions related to each specialized field.

ChatGPT is the acronym for Chat Generative Pre-Trained Transformer, now in its fourth version. Its manufacturer OpenAI (https://youtu.be/--khbXchTeE) has uploaded several promotional videos onto the Internet. This company recently received billions of dollars in investment from Microsoft to integrate the functionalities ChatGPT offers on its platforms. Some edge browser users will have noticed they can already interact with this chatbot as an alternative to a search engine, with the advantage that it answers our questions with synthesized information using natural language. Similar chatbots (such as Google's Bard and Meta's LLaMa) are being integrated into the founder's proprietary platforms and will soon be the main way we interact with the Internet.

The black box of this new generation of chatbots contains Large Language Models (LLMs). LLMs are linguistic models based on a neural network architecture. They are trained to parse sentences (identify the subject, predicate, verb, and inflections) so they are not only able to "understand" a question written in natural language but can also form an answer by combining the words according to a probabilistic model that guides the word or combination of words that is likely to follow according to the training data. This, combined with the ability to access all the information available on the Internet, means that not only do they produce a naturally formed phrase (in other words, one that seems to be produced by a human being) but also that this phrase contains the answer to our question.

Unlike previous efforts, which involved creating a decision tree based on the response to a particular symptom (for example, if the patient responded affirmatively to the question "Does your head hurt?" the algorithm ruled out or branched possible diagnoses), tools like ChatGPT do not depend on a programmed decision structure, and instead emerge from the statistical patterns generated from automated learning. For example, when I asked ChatGPT, "Which would be the more likely psychiatric diagnosis out of Attention Deficit Hyperactivity Disorder and Autism Spectrum Disorder if the patient tends to move around a lot" (and forced its hand after a first interaction asking it to "Choose one of the two disorders to answer my question"), it answered the following, "…the disorder that might be more closely related to this symptom would be ADHD (Attention Deficit Hyperactivity Disorder)...". I think it is difficult not to be surprised by a result like this. This information is obviously already available on the Internet, but ChatGPT was not only able to obtain it, but also to summarize it and highlight the key aspects to make a diagnosis.

The ethical and practical issues raised by tools like this certainly invite discussion. One premise of computing endeavors over the past two decades has been to "move fast and break things" to encourage disruptive innovation. The effects of this slogan have been unfortunate on many levels (Taplin, 2017), but here we would like to focus on those concerning medical practice, particularly psychiatric diagnosis. With the caveat that I am neither a doctor nor a psychiatrist, artificial intelligence (AI) is a topic that has always intrigued me, and I wanted to use this text to discuss ethical or practical issues that matter to me, without attempting to provide an exhaustive list. I believe these reflections are crucial given the revolution in AI tools that is about to take place.

The first important point is the responsibility of providing information to a user. Obviously, anyone can browse the Internet (or a book for that matter) and misuse it. But interacting with a chatbot that responds naturally creates an illusion that the information is accurate and provided by an expert. Who in the AI construction chain will assume the consequences of a misdiagnosis, particularly one made by an AI service company designed to provide diagnoses? Although the information ChatGPT now produces is constrained by regulatory requirements to prevent lawsuits, there are undoubtedly competitors with laxer ethical guidelines (Mantello & Ho, 2023). This brings us to a second point: how accurate can the information it provides be? The example cited by Wei et al. (2023) shows that doctors also make mistakes. However, ChatGPT errors are not necessarily due to a lack of judgment, ignorance, or an inability to cope with complexity, but can instead be caused by inherent biases in the databases used by the chatbot to obtain its information. These biases have already been illustrated in Internet content, in which English is by far the most commonly used language. It is also true that the prevalence, symptoms, and diagnostic comorbidities of particular populations such as those in the global north have been more widely documented. What biases will a chatbot display when a person from an indigenous community, for which there are fewer statistical references, requests a diagnosis?

But not everything is necessarily negative. Chatbots like ChatGPT can provide doctors serving remote locations with different diagnoses for comparative purpose (and not so remote ones, such as doctors in primary and secondary health care dealing with difficult cases). A differential diagnosis performed by artificial intelligence could go some way towards solving this problem with chatbots specifically created and trained for this purpose. They could be a viable alternative, helping clinicians achieve faster, more accurate diagnoses. Although currently limited to text, they could eventually incorporate information such as tone of voice, facial gestures or movement that could make diagnoses extremely accurate. At present, many places have insufficient human capital to provide a timely diagnosis for all the cases existing in Mexico. In this respect, a differential diagnosis performed by artificial intelligence could partly solve the problem.

Ultimately, it is essential to recall that technology must always be at the service of humanity and that its goal is to deal with its problems rather than to create confusion and offer false promises and risky solutions.

REFERENCES

Mantello, P., & Ho, M. T. (2023). Losing the information war to adversarial AI. *AI & SOCIETY*, 1-3. doi: 10.1007/s00146-023-01674-5

Taplin, J. (2017). *Move fast and break things: How Facebook, Google, and Amazon have Cornered Culture and Undermined Democracy*. Pan Macmillan.

Wei, Q., Cui, Y., Wei, B., Cheng, Q., & Xu, X. (2023). Evaluating the performance of ChatGPT in differential diagnosis of neurodevelopmental disorders: A pediatricians-machine comparison. *Psychiatry Research*, *327*, 115351. doi: 10.1016/j.psychres.2023.115351