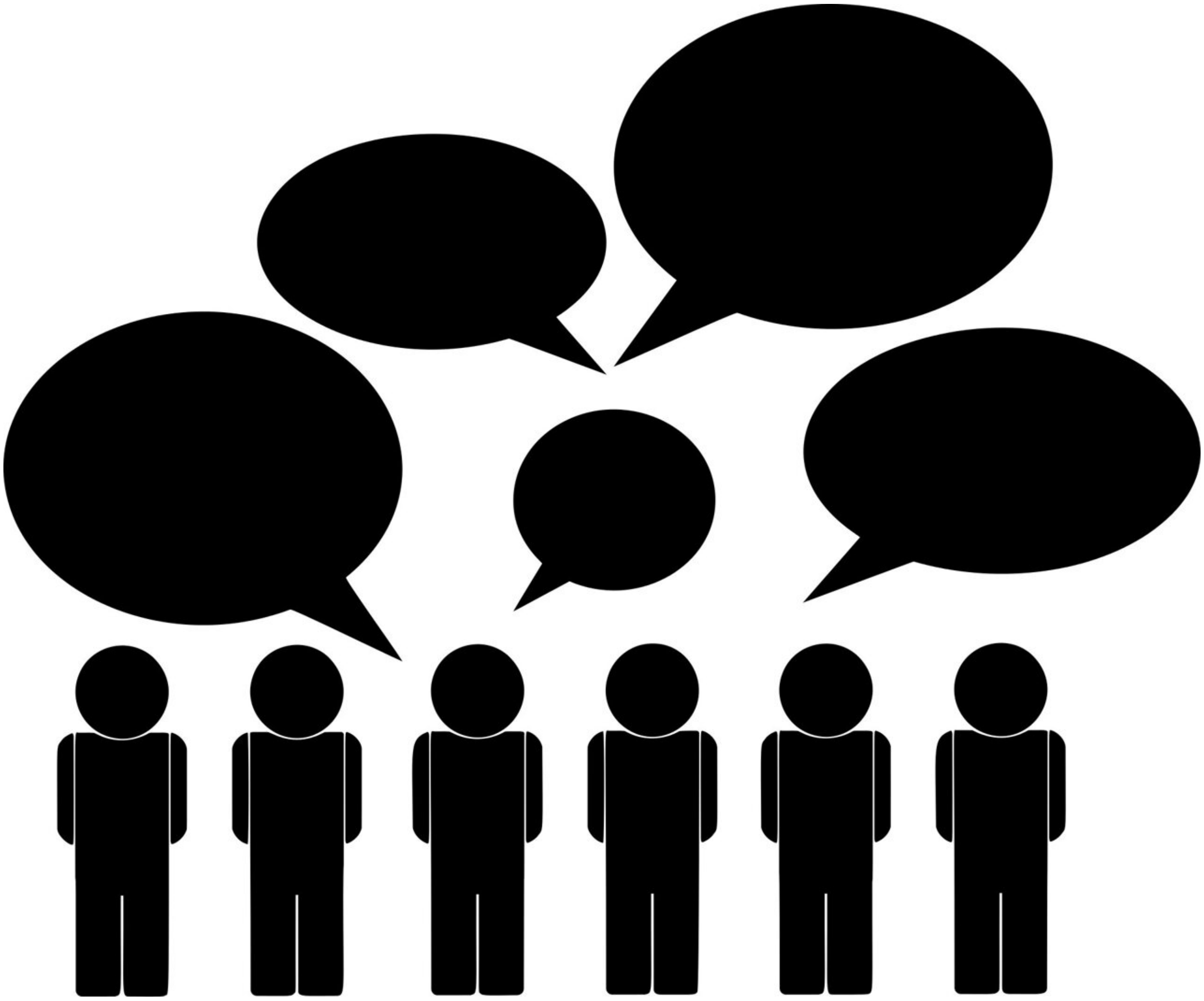
Medical Xpress

AI chatbots work by predicting the next word, so do our brains. Is there a connection?

May 24 2023, by Taylor McNeil



Credit: Pixabay/CC0 Public Domain

When we hear or read a string of words, we are unconsciously making predictions about what's coming next. That's much like what AI chatbots like ChatGPT and Bard do. At root they are predicting what the next word should be in their responses, with human-like accuracy.

Large language model AIs and human brains are fundamentally different, but that doesn't mean that AIs don't have a role in understanding the human brain. Gina Kuperberg, the Dennett Stibel Professor of Cognitive Science and professor of psychology, studies language comprehension—how the brain makes sense of language. She and her colleagues use AI to help understand human cognition.

"As I'm speaking to you right now, you're taking every single word and you're incrementally building up meaning from the combination of words that I'm using," she told me. "You're doing that really quickly. You're not waiting until the end of each sentence or each clause to do that. You're doing that as you're going along, and based on your interpretation, you're predicting what I'm going to say next."

In her research, she is trying to figure out the neurobiology of that mechanism, where in the brain it takes place, how quickly it happens, and precisely how it happens.

Tufts Now spoke with Kuperberg, who is also a psychiatrist in the Department of Psychiatry at Massachusetts General Hospital, to learn more about language and learning.

Tufts Now: What similarities and differences are there between large language model AI chatbots and humans in terms of learning languages?

Gina Kuperberg: A large language model is trained with very, very large bodies of text. It's trained to predict upcoming words. In doing so, it picks up the statistical contingencies, the very complex higher-order relationships between words.

Our brains are constantly predicting the next word, too, and our brains are very good at picking up statistical contingencies and very complex higher order relationships between words.

So in one sense, you could say the reason why these large language models are so successful is because they capitalize on something that our brains do all the time. But just because these chatbots can predict the next word and perform so well doesn't necessarily mean that the precise mechanisms by which they do that are the same as those used by the human brain.

Is there anything that cognitive neuroscientists can learn from these AIs to understand the human brain?

Well, it may be that they are learning similar representations to the human brain, and that may tell us something. But it's also important to recognize that, even though these large language models produce language so well, the precise computational mechanism by which they get there are quite different from those used by the human brain.

We know quite a lot about the neurobiology of cognitive processes, including language in the brain. And we also know quite a lot about how these large language models are constructed. And they're very different.

For example, in the human brain, with different layers of cortical structure, we know that there's constant interaction and feedforward and feedback interactions between these different layers.

And we know that the state-of-the-art language models like ChatGPT, the transformer-based language models, don't have feedback interactions between layers. They're wired differently from the human brain.

So just because you can get a really good end result with them—and we can use that to try and understand human language comprehension and production—that doesn't that we can make claims about the precise connectivity and the precise mechanisms by which it reaches that end goal.

Do you use these AIs in your research?

We use them more as a tool to try and understand how the brain works. For example, we know that the brain predicts upcoming words, so we use ChatGPT to predict upcoming words and then look to see how the predictability of each upcoming word correlates with brain activity.

For another project in my lab, we're looking at produced speech, and we're looking to see the similarities between ChatGPT predictions of each word that someone produces and what the brain produces.

We're also seeing whether the predictability of each word produced by people with schizophrenia is reduced compared to our controls. And then we're also giving the language model progressively less context to see whether that makes a difference.

I'm curious about the work that you're doing with schizophrenia—can you explain that a bit more?

Speech produced by people with schizophrenia, which is a severe mental disorder, can in some patients be characterized by what's known as thought disorder. That means that their language can be disorganized and difficult to understand.

We're trying to figure out whether we can describe the speech produced by people with schizophrenia in terms of reduced predictability, and whether people with schizophrenia predict on the basis of global or more local context. A large language model is ideal for testing this hypothesis. My graduate student Tori Sharpe is leading the project.

We have access to speech samples produced by people with schizophrenia in their very first episode of psychosis as they're describing pictures, and we can assess the predictability of each word that they produce. We find that predictability of each word is reduced in schizophrenia.

That's interesting, but what's more interesting is that, based on only the local context, people with schizophrenia produce language that is just as predictable as the language produced by people without schizophrenia. The way that we were able to figure this out is by giving the large language model different amounts of context to make its prediction.

When you give the model words 1–50, and estimate the predictability of word 51, and then compare people with and without schizophrenia, you find that the people with schizophrenia produce less predictable words. But if you give the model words 42 to 49, and use this limited context to estimate predictability, then it turns out that there is no difference between the predictability of the people with schizophrenia and healthy controls.

That's important, because it tells us that people with schizophrenia are able to predict the next word based on local rather than global context. That's been a hypothesis in the research for a while, but now with large language models, we can systematically test it.

Is that related to memory—that people with schizophrenia can't hold words one through 41 in short-term memory, only words 42 to 49?

It may well be related to their working memory, but that's still an open question. In our current research with one measure of short-term working memory, we saw no difference. It may be a more specific kind of verbal working memory that is impaired in patients. We also found that the local versus global difference predicted the degree of thought disorder within the patient group, which is important because it provides a way to operationalize this clinical symptom.

Are there other uses of AIs in your research?

A grad student in my lab, Samer Nour Eddine, is developing a "toy" language model based on the principles of an algorithm known as predictive coding. The nice thing about toy language models is that, unlike large language models where the inner workings are somewhat obscure, you can understand how they work and tweak their precise mechanisms and algorithms.

The predictive coding algorithm has been very successful in explaining important phenomena in the human visual system. What Samer did was to use the precise architecture and algorithm of predictive coding to simulate a neural signature called the N400, which is highly sensitive to the predictability of each incoming word. When you get an incoming word that is unpredictable, the N400 is large, but it gets smaller and smaller as words become more and more predictable.

Samer has used predictive coding to simulate this N400 effect. There's no way a large language model can do that. A large language model can describe the sensitivity of this N400 effect to predictability, but it can't simulate its time course—the rise and the fall of the N400 over time—and it can't simulate other factors that the N400 is sensitive to.

What's cool about predictive coding is that it's biologically and cognitively plausible—it was first used to describe phenomenon and the visual system. It has lots of feedforward and feedback connections across layers just like the human brain. I think some people are trying to develop large language models based on predictive coding principles, but it is quite different from the state-of-the-art things like ChatGPT, whose architecture is quite different from a predictive coding architecture.

Some people are alarmed by chatbots' abilities. You seem to be using them in useful ways in your research. What is your general take?

It's just like anything, isn't it? When paper first came into use, Plato was alarmed; he thought people would use it unwisely and it would destroy the brain's ability to think and use memory.

I think one should certainly be alarmed in terms of large language models being unprecedented tools for disseminating misinformation. But there are always amazing uses for new technology and new AI tools.

It's going to change the way people write. It can't change the way people think. It's not going to make us obsolete. So, like the internet, there are really good things to be done with AI tools, and really bad things, and it's up to us to use these things wisely and carefully.

Provided by Tufts University

Citation: AI chatbots work by predicting the next word, so do our brains. Is there a connection?

(2023, May 24) retrieved 6 May 2024 from https://medicalxpress.com/news/2023-05-ai-chatbots-word-brains.html