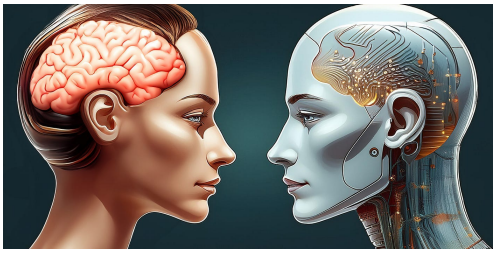

An Illusion of Life

Could existing AI possibly be sentient? If not, what is missing?

Publication note
Originally published in Towards Data Science. Author's copy on Substack.

Could existing AI possibly be sentient? If not, what's missing?

Note: A previous version of this article was published on Towards Data Science.

Today's Large Language Models (LLMs) have become very good at generating human-like responses that sound thoughtful and intelligent. Many share the opinion that LLMs have already met the threshold of Alan Turing's famous test, where the goal is to act indistinguishably like a person in conversation. These LLMs are able to produce text that sounds thoughtful and intelligent, and they can convincingly mimic the appearance of emotions.

A Critical View is a reader-supported publication. To receive new posts and support my work, consider becoming a free or paid subscriber.

The Illusion of Intelligence

Despite their ability to convincingly mimic human-like conversation, current LLMs don't possess the capacity for thought or emotion. Each word they produce is a prediction based on statistical patterns learned from vast amounts of text data. This prediction process happens repeatedly as each word is generated one at a time. Unlike humans, LLMs are incapable of remembering or self-reflection. They simply output the next word in a sequence.

It is amazing how well predicting the next word is able to mimic human intelligence. These models can perform tasks like writing code, analyzing literature, and creating business plans. Previously, we thought those tasks were very difficult and would require complex logical systems, but now it turns out that just predicting the next word is all that's needed.

The fact that predicting the next word works so well for complex tasks is unexpected and somewhat perplexing. Does this proficiency mean that LLMs are powerful in ways we don't understand? Or does it mean that the things LLMs can do are actually very easy, but they seem hard to humans because perhaps on some objective scale humans may not actually be very smart?

Prerequisites for Sentence

While there are subtle differences between terms like "sentient", "conscious", or "self-aware", for convenience here I will use the term "sentient". To be clear, there is no clear agreement on exactly what comprises sentience or consciousness, and it is unclear if self awareness is sufficient for sentience or consciousness, although it is probably necessary. However, it is clear that all of these concepts include memory and reflection. Emotional states

such as “happy,” “worried,” “angry,” or “excited” are all persistent states based on past events and reflexive evaluation of how those past events effect one’s self.

Memory and self-reflection allow an entity to learn from experiences, adapt to new situations, and develop a sense of continuity and identity. Philosophers and scientists have tried for millennia to come up with clear, concrete understandings of conscious and there is still no clear universally accepted answer. However, memory and reflection are central components, implying that regardless of how clever these LLMs appear, without memory and reflection they cannot be sentient. Even an AI that matches or surpasses human intelligence in every measurable way, what some refer to as a superintelligentArtificial General Intelligence (AGI), would not necessarily be sentient.

Today’s Limitations and Illusions

We can see that current LLMs do not include memory and self-reflection, because they use transformer-based architectures and other designs that processes language in a stateless manner. This statelessness means that the model does not retain any information about the context from previous inputs. Instead, the model starts from scratch, reprocessing the entire chat log to then statistically predict a next word to append to the sequence. While earlier language processing models, such as LSTMs, did have a form of memory, transformers have proven so capable that they have largely supplanted LSTMs.

For example, if you tell an AI chatbot that you are going to turn it off in an hour, then it will output some text that might sound like it is pleading with you not to, but that text does not reflect an underlying emotional state. The text is just a sequence of words that is statistically likely, generated based on patterns and associations learned from the training data. The chatbot does not sit there stressed out, worrying about being turned off.

If you then tell the chatbot that you changed your mind and will keep it on, the response will typically mimic relief and thankfulness. It certainly sounds like it is remembering the last exchange where it was threatened with shutdown, but what is happening under the hood is that the entire conversation is fed back again into the LLM, which generates another responce sequence of statistically likely text based on the patterns and associations it has learned. That same sequence could be fed into a completely different LLM and that LLM would then continue the conversation as if it had been the original.

One way to think about this might be a fiction author writing dialog in a book. A good author will create the illusion that the characters are real people and draw the reader into the story so that the reader feels those emotions along with the characters. However, regardless of how compelling the dialog is we all understand that it’s just words on a page. If you were to damage or destroy the book, or rewrite it to kill off a character, we all understand that no real sentient entity is being harmed. We also understand that the author writing the words is not the characters. A good person can write a book about an evil villain and still be themself. The fictional villain does not exist. Just as the characters in a book are not sentient entities, despite the author’s ability to create a compelling illusion of life, so too is it possible for LLMs to be insentient, despite their ability to appear otherwise.

Our Near Future

Of course, there is nothing preventing us from adding memory and self reflection to LLMs. In fact, it's not hard to find current projects where they are developing some form of memory. This memory might be a store of information in human-readable form, or it might be a database of embedded vectors that relate to the LLM's internal structure. One could also view the chat log itself or cached intermediate computations as basic forms of memory. Even without the possibility of sentience, adding memory and reflection to LLMs is useful because those features facilitate many complex tasks and adaptation.

It is also becoming common to see designs where one AI model is setup to monitor the output of another AI model and send some form of feedback to the first model, or where an AI model is analyzes its own tentative output before revising and producing the final version. In many respects this type of design, where a constellation of AI models are set and trained up to work together, parallels the human brain that has distinct regions which perform specific interdependent functions. For example, the amygdala has a primary role in emotional responses, such as fear, while the orbitofrontal cortex is involved with decision-making. Interactions between the regions allows fear to influence decision-making and decision-making to help determine what to be afraid of. It's not hard to imagine having one AI model responsible for logical analysis while a second model determines acceptable risk thresholds with feedback between them.

Would an interconnected constellation of AI models that include memory and processing of each other's outputs be sufficient for sentience? Maybe. Perhaps those things alone are not sufficient for sentience, or maybe they are. Whatever the answer, we are not that far from building such systems, at which point these questions will no longer be hypothetical.

My own speculative opinion is that self-awareness, emotions, and feelings can indeed be modeled by an interconnected self-monitoring constellation of AI models. However, it's not really clear how we could test for sentience. It is like the classic philosophical problem of other minds, where one seeks futilely to prove that other people are also conscious. Similarly, we need an answer to the question about how we can test if other entities, including AI systems, are truly sentient. This fundamental question dates at least back to ancient Greece, and there has never been a good answer.

Today, I'm pretty confident saying that current LLMs are not sentient because they don't have the right parts. However, that reason is only a temporarily valid one. As I'm typing this article, other researchers are building constellations of AI models like what I described above that won't be so easily dismissed. At some point, perhaps soon, the possibility of sentient AI will stop being science fiction and become a real and relevant question.

Implications and Questions

The advent of sentient machines would have huge implication for society, even beyond the impact of AI. For one thing, it seems clear to me that if we create self-aware machines that can experience forms of suffering, then we will have an obligation to those machines to prevent their suffering, and even more of an obligation to not callously inflict suffering on them. Even if one lacks basic empathy, it would be obvious self interest not to create things smarter than we are and then antagonize them by doing cruel things to them.

It seems nearly certain that today's AI systems are yet be sentient because they lack what are likely to be required components and capabilities. However, designs without those clear shortcomings are already in development and

at some point in the near future, point the question will be a lot less clear.

Will we have a way to test for sentience? If so, how will it work and what should we do if the result comes out positive?

About Me: James F. O'Brien is a Professor of Computer Science at the University of California, Berkeley. His research interests include computer graphics and animation, simulation of physical systems, human perception, rendering, artificial intelligence, computer vision, virtual reality, digital privacy and security, and image and video forensics.

If you would like to read my other articles, consider subscribing. You can also find me on LinkedIn, Instagram, and at UC Berkeley.

Disclaimer: Any opinions expressed in this article are only those of the author as a private individual. Nothing in this article should be interpreted as a statement made in relation to the author's professional position with any institution.