UC SANTA BARBARA

THE Current

November 1, 2023

Harrison Tasoff

# The mind of the machine

In a startlingly short time span, artificial intelligence has evolved from an academic undertaking into a practical tool. Visual models like DALL·E can create images in any style an individual might fancy, while large language models (LLMs) like Chat GPT can generate essays, write computer code and suggest travel itineraries. When prompted, they can even correct their own mistakes.

As AI models become ever more sophisticated and ubiquitous, it's crucial to understand just what these entities are, what they can do and how they think. These models are becoming very similar to humans, and yet they are so very different from us. This unique combination makes AI intriguing to contemplate.

For instance, large AI models are trained on immense amounts of information. But it isn't clear to what extent they understand this data as a coherent system of knowledge. UC Santa Barbara's Fabian Offert explores this idea in a short article featured in the anthology "ChatGPT und andere Quatschmaschinen" (Transcript Verlag, 2023), which translates to "ChatGPT and other nonsense machines."

"People have been claiming that the large language models, and Chat GPT in particular, have a so-called 'world model' of certain things, including computation," said Offert, an assistant professor of digital humanities. That is, it's not just superficial knowledge that coding words often appear together, but a more comprehensive understanding of computation itself.

Even a basic computer program can produce convincing text with a Markov chain, a simple algorithm that uses probability to predict the next token in a sequence based on what's come before. The nature of the output depends on the reference text and the size of the token (e.g. a letter, a word or a sentence). With the proper parameters and training source, this can produce natural text mimicking the style of the training sample.

But LLMs display abilities that you wouldn't expect if they were merely predicting the next word in a sequence. For instance, they can produce novel, functional computer code. Formal languages, like computer languages, are much more rigid and well defined than the natural languages that we speak. This makes them more difficult to navigate holistically, because code needs to be completely correct in order to parse; there's no wiggle room. LLMs seem to have contextual memory in a way that simple Markov chains and predictive algorithms don't. And this memory gives rise to some of their novel behaviors, including their ability to write code.

Offert decided to pick Chat GPT's brain by asking it to carry out a few tasks. First, he asked it to code a Markov chain that would generate text based on the novel "Eugene Onegin," by Alexander Pushkin. After a couple false starts, and a bit of coaxing, the AI produced a working Python code for a word-level Markov chain approximation of the book.

Next, he asked it to simply simulate the output of a Markov chain. If Chat GPT truly had a model of computation beyond just statistical prediction, Offert reasoned that it should be able to estimate the output of a program without running it. He found that the AI could simulate a Markov chain at the level of words and phrases. However, it couldn't estimate the output of a Markov chain letter-by-letter. "You should get somewhat coherent letter salad, but you don't," he said.

This outcome struck Offert as rather odd. Chat GPT clearly possessed a more nuanced understanding of programming because it successfully coded a Markov chain during the first task. However, if it truly possessed a concept of computation, then predicting a letter-level Markov chain should be quite easy for it. This requires far less computation, memory and effort than predicting the outcome at the word level, which it was able to do. That said, there are other ways that it could've accomplished the word-level prediction simply because LLMs are, by design, good at generating words.

“Based on this result, I would say Chat GPT does not have a world model of computation,” Offert opined. “It’s not simulating a good old Turing machine with access to the full capabilities of computation.”

Offert’s goal in this paper was merely to raise questions, though, not answer them. He was simply chatting with the program, which isn’t proper methodology for a scientific investigation. It’s subjective, uncontrolled, not reproducible and the program might update from one day to the next. “It’s really more like a qualitative interview than it is a controlled experiment,” he explained. Just probing the black box, if you will.

Offert wants to develop a better understanding of these new entities that have come into being over the last few years. “My interest is really epistemological,” he said. “What can we know with these things? And what can we know about these things?” Of course, these two questions are inextricably linked.

These topics have begun to attract the interests of engineers and computer scientists as well. “More and more, the questions that technical researchers ask about AI are really, at their core, humanities questions,” Offert said. “They’re about fundamental philosophical insights, like what it means to have knowledge about the world and how we represent knowledge about the world.”

This is why Offert believes that the humanities and social sciences have a more active part to play in the development of AI. Their role could be expanded to inform how these systems are developed, how they’re used and how the public engages with them.

The differences between artificial and human intelligences are perhaps even more intriguing than the similarities. “The alien-ness of these systems is actually what is interesting about them,” Offert said. For example, in a previous paper, he revealed that the way AI categorizes and recognizes images can be quite strange from our perspective. “We can have incredibly interesting, complex things with emergent behaviors that are not just machine humans.”

Offert is ultimately trying to understand how these models represent the world and make decisions. Because they do have knowledge about the world, he assures us — connections gleaned from their training data. Going beyond epistemological interest,

the topic is also of practical importance for aligning the motivations of AI with those of its human users.

As tools like Chat GPT become more widely used, they bring formerly unrelated disciplines closer together. For instance, essay writing and noise removal in astronomy are now both connected to the same underlying technology. According to Offert, that means we need to start looking at the technology itself in greater detail as a fundamentally new way of generating knowledge.

With a three-year grant from the Volkswagen Foundation on the topic of AI forensics, Offert is currently exploring machine visual culture. Image models have become so large, and seen so much data, he explained, that they've developed idiosyncrasies based on their training material. As these tools become more widespread, their quirks will begin feeding back into human culture. As a result, Offert believes it's important to understand what's going on under the hood of these AI models.

"It's an exciting time to be doing this work," he said. "I wouldn't have imagined this even five years ago."

Tags
Artificial Intelligence

Media Contact

**Harrison Tasoff**

Science Writer

(805) 893-7220

harrisontasoff@ucsb.edu

---

## About UC Santa Barbara

The University of California, Santa Barbara is a leading research institution that also provides a comprehensive liberal arts learning experience. Our academic community of faculty, students, and staff is characterized by a culture of interdisciplinary collaboration that is responsive to the needs of our multicultural and global society.

All of this takes place within a living and learning environment like no other, as we draw inspiration from the beauty and resources of our extraordinary location at the edge of the Pacific Ocean.