In a neural network, data points are represented by abstract mathematical quantities known as “vectors,” and “the whole process happens in some high-dimensional vector space, which is very hard for people to interpret,” says Azrieli Early Career Faculty Fellow Yonatan Belinkov. “It’s not a two-dimensional or even three-dimensional space that we can understand. It’s maybe 1,000 dimensions. We don’t really know what’s going on there . . . it’s really a black box.”

Chatbots Say the Darndest Things

By Dan Falk
Photographs by Boaz Perlstein

Neural networks are becoming more human-like and powerful, but we still don’t really understand how they work.

What’s it like to specialize in a branch of science that’s in the news almost every day? That’s the unusual position in which Yonatan Belinkov finds himself.

He’s an expert on artificial intelligence (AI) and natural language processing, which includes the study of large language models (LLMs) such as ChatGPT. Released last November by OpenAI, ChatGPT is a sophisticated chatbot that can generate text in response to just about any prompt a user gives it. The text it produces can seem very human.

But the technology has also proven controversial. On one hand, ChatGPT appears to be bolstering productivity in fields such as marketing, grant writing and data analysis. At the same time, there are concerns about its effects in schools and universities — it can produce passable undergraduate-level essays in the blink of an eye — and worries that it threatens journalism and even democracy, with its potential to flood the world with fake news.

That’s because sentences produced by ChatGPT aren’t necessarily true. Belinkov, a professor of computer science and Azrieli Early Career Faculty Fellow at Technion–Israel Institute of Technology, put the software to the test recently by asking it who won the Nobel Peace Prize in 1948.

“It told me that the United Nation Committee on Civil Rights won the Nobel Prize,” he says, “and it told me why it won the Nobel Prize. It gave a very convincing answer — except that it’s false. The prize wasn’t awarded that year. Gandhi was nominated, but he was murdered just a few days before the decision, so it wasn’t awarded. But ChatGPT was very convinced.” (Cases where chatbots seem to go off the rails in this manner have been dubbed “hallucinations” and are thought to be triggered when the system strays too far from its training data.)

Falsehood is only one problem. AI language systems have also been known to perpetuate biases. For example, studies have shown that when prompted with statements such as “the nurse said that . . .” the system is more likely to complete the sentence on the assumption that the nurse is a woman rather than a man. And it found the reverse bias when prompted with the word “doctor.”

These are big changes from a decade ago, when chatbots could barely string together a coherent sentence. “In 2012, when I started my PhD, I couldn’t imagine anything like what we have today,” says Belinkov. “It’s a little challenging to be in a field that is very, very hot, because progress is so fast. Every day, there’s a new research article that I need to read. Staying up to date can be tricky.”

The key development enabling this leap forward is the rise of “deep learning” architectures

known as artificial neural networks. These networks use a series of "layers" of mathematical processing to assess the information they're fed. The connections between the layers are assigned weights that reflect the importance of each connection relative to the others, and those weights are adjusted as the network is exposed to more and more input data. Finally, the last layer produces an output. In recent years, neural networks have become proficient at recognizing faces, translating languages and, with programs such as ChatGPT, creating human-like text. (A few months ago, an even newer version, called GPT-4, was released; it can create websites in minutes, explain jokes and suggest recipes based on a photo of what's in your fridge.)

Although programs like ChatGPT are certainly impressive, shortcomings such as bias and hallucinations demand our attention, Belinkov says. But exactly why a neural network gives one particular output rather than another has always been rather opaque. While the overall architecture of such networks is well understood, the actual cause-and-effect links from input to output can seem deeply mysterious. In a neural net, data points are represented by abstract mathematical quantities known as "vectors," and "the whole process happens in some high-dimensional vector space, which is very hard for people to interpret," Belinkov says. "It's not a two-dimensional or even three-dimensional space that we can understand. It's maybe 1,000 dimensions." As a result, he adds, "we don't really know what's going on there. In this sense, it's really a black box."

With no way to track in detail what that vast array of vectors is actually doing, how can one tackle problems like bias in chatbots and similar AI systems? Belinkov's strategy is to probe the network's internal structure, examining which parts of a network are active at each stage of the process. He describes the process as an "intervention" in which one runs the program multiple times, each time tweaking one, or a few, components. "And every once in a while, you find there's a neuron" — he uses the same term used to refer to cells in the brain — "where, when you switch it off, the machine gets lost. And that tells you, OK, maybe that's where things are happening."

Belinkov draws a comparison to an old-time arcade game in which a ball is inserted at the top of a wide, flat box with an array of pegs in it that allow the ball to move along some paths but not others. The user can remove or insert various pegs in order to try to steer the ball one way or another. Often, he says, many of the pegs — or in the case of a chatbot, network connections — don't actually affect the end result.

"We have to understand where, in the system, there's a switch — maybe one or more switches — that determines what it will say," explains Belinkov. Only a few of these switches would actually determine what text the system produces in response to a particular input. "So, this question of 'why' is, I think, very much related to the question of 'where,'" he says. "If we can pinpoint one switch, or multiple switches, that are responsible for producing a particular output, then we have an answer of why."

The fact that some parts of the network seem to be much more important than others comes as a surprise, Belinkov says, given how interconnected the whole setup is. "If everything is connected to everything, then computation should be distributed and information should be likewise distributed. But it turns out that it's not. For certain inputs, certain neurons activate much more than everything else."

Another potential strategy is to feed the system new inputs, but that's not as easy as it sounds, cautions Belinkov. For example, if an LLM was trained two years ago, some of its outputs will be out of date — but retraining it on the entire internet could cost millions of dollars. So, a more efficient strategy is to just tweak the specific parts of the system that have been implicated in producing problematic outputs, Belinkov explains. To combat a problem like gender bias, then, a better approach, he says, is "to look at the internal structure, identify the roles of different components, and then change them or edit them in a way that we think is desirable."

Although programs like ChatGPT are certainly impressive, shortcomings such as bias and hallucinations demand our attention, Belinkov says. But exactly why a neural network gives one particular output rather than another has always been rather opaque. While the overall architecture of such networks is well understood, the actual cause-and-effect links from input to output can seem deeply mysterious.

David Bau, a computer scientist at Northeastern University in Boston who has collaborated with Belinkov, notes that the remarkable capabilities of large language models "have been a genuine surprise, even among the experts of the field." Belinkov's research is important, he says, because the opacity of LLMs makes traditional debugging techniques ineffective. "Yonatan's research about these models' internal mechanisms will be critical in closing our gaps in understanding. Already, his work is laying the groundwork to help identify and solve potential problems of bias, robustness, misinformation and privacy that can emerge in large models."

Studying the internal architecture of AI systems may sound like an esoteric pursuit, but with AI systems playing an ever-increasing role in society, the stakes are high. When computers were first invented some 75 years ago, the assumption was that they would help us. And in many ways they have. But to continue to aid humankind, the objectives of AI systems have to be aligned with our own. Indeed, Australian philosopher Toby Ord has estimated there's a one-in-ten chance that "non-aligned" AI will trigger a catastrophe during the next century.

Belinkov admits that for much of his career he considered the "alignment" problem to be a low priority, something to worry about in the far future. The threat "seemed so far-fetched and so remote," he says. "But I have to say that, recently, I've started thinking that, yes, maybe it is time for both scientists and legislators to start thinking about the alignment problem more seriously. We need to make sure that AI systems are not biased and perform what we want them to perform. But at the same time, we should think about longer-term risks." ▲●■

Studying the internal architecture of AI systems may sound like an esoteric pursuit, but with these systems playing an ever-increasing role in society, the stakes are very high. "We need to make sure that AI systems are not biased and perform what we want them to perform," says Belinkov. "At the same time, we should think about longer-term risks."