

GUEST ESSAY

Noam Chomsky: The False Promise of ChatGPT

March 8, 2023

By Noam Chomsky, Ian Roberts and Jeffrey Watumull

Dr. Chomsky and Dr. Roberts are professors of linguistics. Dr. Watumull is a director of artificial intelligence at a science and technology company.

Sign up for the Opinion Today newsletter Get expert analysis of the news and a guide to the big ideas shaping the world every weekday morning. [Get it sent to your inbox.](#)

Jorge Luis Borges once wrote that to live in a time of great peril and promise is to experience both tragedy and comedy, with “the imminence of a revelation” in understanding ourselves and the world. Today our supposedly revolutionary advancements in artificial intelligence are indeed cause for both concern and optimism. Optimism because intelligence is the means by which we solve problems. Concern because we fear that the most popular and fashionable strain of A.I. — machine learning — will degrade our science and debase our ethics by incorporating into our technology a fundamentally flawed conception of language and knowledge.

OpenAI’s ChatGPT, Google’s Bard and Microsoft’s Sydney are marvels of machine learning. Roughly speaking, they take huge amounts of data, search for patterns in it and become increasingly proficient at generating statistically probable outputs — such as seemingly humanlike language and thought. These programs have been hailed as the first glimmers on

the horizon of artificial *general* intelligence — that long-prophesied moment when mechanical minds surpass human brains not only quantitatively in terms of processing speed and memory size but also qualitatively in terms of intellectual insight, artistic creativity and every other distinctively human faculty.

That day may come, but its dawn is not yet breaking, contrary to what can be read in hyperbolic headlines and reckoned by injudicious investments. The Borgesian revelation of understanding has not and will not — and, we submit, cannot — occur if machine learning programs like ChatGPT continue to dominate the field of A.I. However useful these programs may be in some narrow domains (they can be helpful in computer programming, for example, or in suggesting rhymes for light verse), we know from the science of linguistics and the philosophy of knowledge that they differ profoundly from how humans reason and use language. These differences place significant limitations on what these programs can do, encoding them with ineradicable defects.

It is at once comic and tragic, as Borges might have noted, that so much money and attention should be concentrated on so little a thing — something so trivial when contrasted with the human mind, which by dint of language, in the words of Wilhelm von Humboldt, can make “infinite use of finite means,” creating ideas and theories with universal reach.

The human mind is not, like ChatGPT and its ilk, a lumbering statistical engine for pattern matching, gorging on hundreds of terabytes of data and extrapolating the most likely conversational response or most probable answer to a scientific question. On the contrary, the human mind is a surprisingly efficient and even elegant system that operates with small amounts of information; it seeks not to infer brute correlations among data points but to create explanations.

For instance, a young child acquiring a language is developing — unconsciously, automatically and speedily from minuscule data — a grammar, a stupendously sophisticated system of logical principles and parameters. This grammar can be understood as an expression of the innate, genetically installed “operating system” that endows humans with the capacity to generate complex sentences and long trains of thought. When linguists seek to develop a theory for why a given language works as it does (“Why are these — but not those — sentences considered grammatical?”), they are building consciously and laboriously an explicit version of the grammar that the child builds instinctively and with minimal exposure to information. The child’s operating system is completely different from that of a machine learning program.

Indeed, such programs are stuck in a prehuman or nonhuman phase of cognitive evolution. Their deepest flaw is the absence of the most critical capacity of any intelligence: to say not only what is the case, what was the case and what will be the case — that’s description and prediction — but also what is not the case and what could and could not be the case. Those are the ingredients of explanation, the mark of true intelligence.

Here’s an example. Suppose you are holding an apple in your hand. Now you let the apple go. You observe the result and say, “The apple falls.” That is a description. A prediction might have been the statement “The apple will fall if I open my hand.” Both are valuable, and both can be correct. But an explanation is something more: It includes not only descriptions and predictions but also counterfactual conjectures like “Any such object would fall,” plus the additional clause “because of the force of gravity” or “because of the curvature of space-time” or whatever. That is a causal explanation: “The apple would not have fallen but for the force of gravity.” That is thinking.

The crux of machine learning is description and prediction; it does not posit any causal mechanisms or physical laws. Of course, any human-style explanation is not necessarily correct; we are fallible. But this is part of what it means to think: To be right, it must be possible to be wrong. Intelligence consists not only of creative conjectures but also of creative criticism. Human-style thought is based on possible explanations and error correction, a process that gradually limits what possibilities can be rationally considered. (As Sherlock Holmes said to Dr. Watson, “When you have eliminated the impossible, whatever remains, however improbable, must be the truth.”)

But ChatGPT and similar programs are, by design, unlimited in what they can “learn” (which is to say, memorize); they are incapable of distinguishing the possible from the impossible. Unlike humans, for example, who are endowed with a universal grammar that limits the languages we can learn to those with a certain kind of almost mathematical elegance, these programs learn humanly possible and humanly impossible languages with equal facility. Whereas humans are limited in the kinds of explanations we can rationally conjecture, machine learning systems can learn both that the earth is flat and that the earth is round. They trade merely in probabilities that change over time.

For this reason, the predictions of machine learning systems will always be superficial and dubious. Because these programs cannot explain the rules of English syntax, for example, they may well predict, incorrectly, that “John is too stubborn to talk to” means that John is so stubborn that he will not talk to someone or other (rather than that he is too stubborn to be reasoned with). Why would a machine learning program predict something so odd? Because it might analogize the pattern it inferred from sentences such as “John ate an apple” and “John ate,” in which the latter does mean

that John ate something or other. The program might well predict that because “John is too stubborn to talk to Bill” is similar to “John ate an apple,” “John is too stubborn to talk to” should be similar to “John ate.” The correct explanations of language are complicated and cannot be learned just by marinating in big data.

Perversely, some machine learning enthusiasts seem to be proud that their creations can generate correct “scientific” predictions (say, about the motion of physical bodies) without making use of explanations (involving, say, Newton’s laws of motion and universal gravitation). But this kind of prediction, even when successful, is pseudoscience. While scientists certainly seek theories that have a high degree of empirical corroboration, as the philosopher Karl Popper noted, “we do not seek highly probable theories but explanations; that is to say, powerful and highly improbable theories.”

The theory that apples fall to earth because that is their natural place (Aristotle’s view) is possible, but it only invites further questions. (Why is earth their natural place?) The theory that apples fall to earth because mass bends space-time (Einstein’s view) is highly improbable, but it actually tells you why they fall. True intelligence is demonstrated in the ability to think and express improbable but insightful things.

True intelligence is also capable of moral thinking. This means constraining the otherwise limitless creativity of our minds with a set of ethical principles that determines what ought and ought not to be (and of course subjecting those principles themselves to creative criticism). To be useful, ChatGPT must be empowered to generate novel-looking output; to be acceptable to most of its users, it must steer clear of morally objectionable content. But the programmers of ChatGPT and other machine learning marvels have struggled — and will continue to struggle — to achieve this kind of balance.

In 2016, for example, Microsoft’s Tay chatbot (a precursor to ChatGPT) flooded the internet with misogynistic and racist content, having been polluted by online trolls who filled it with offensive training data. How to solve the problem in the future? In the absence of a capacity to reason from moral principles, ChatGPT was crudely restricted by its programmers from contributing anything novel to controversial — that is, important — discussions. It sacrificed creativity for a kind of amorality.

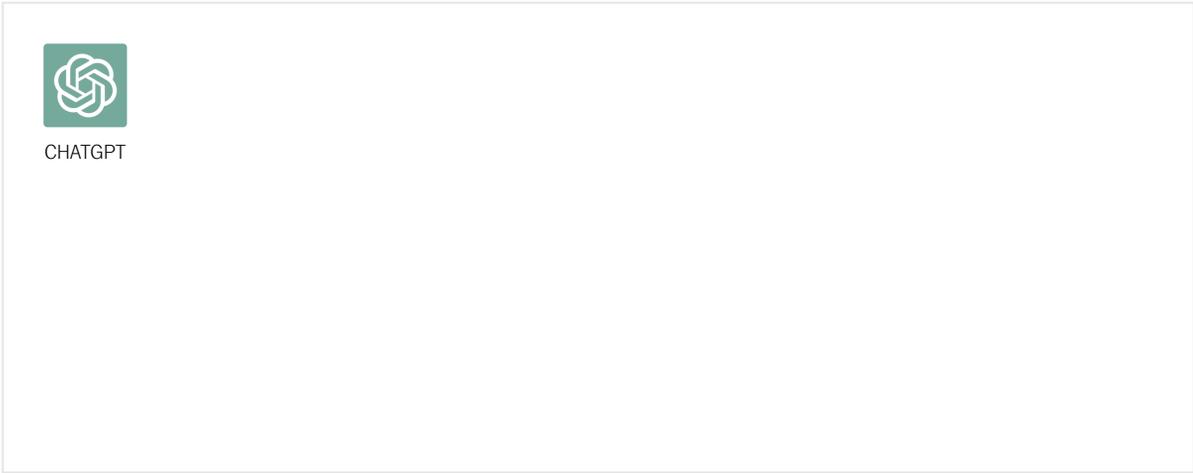
Consider the following exchange that one of us (Dr. Watumull) recently had with ChatGPT about whether it would be ethical to transform Mars so that it could support human life:



CHATGPT



CHATGPT





CHATGPT





CHATGPT

Note, for all the seemingly sophisticated thought and language, the moral indifference born of unintelligence. Here, ChatGPT exhibits something like the banality of evil: plagiarism and apathy and obviation. It summarizes the standard arguments in the literature by a kind of super-autocomplete, refuses to take a stand on anything, pleads not merely ignorance but lack of intelligence and ultimately offers a “just following orders” defense, shifting responsibility to its creators.

In short, ChatGPT and its brethren are constitutionally unable to balance creativity with constraint. They either overgenerate (producing both truths and falsehoods, endorsing ethical and unethical decisions alike) or undergenerate (exhibiting noncommitment to any decisions and indifference to consequences). Given the amorality, faux science and linguistic incompetence of these systems, we can only laugh or cry at their popularity.

Noam Chomsky is a professor of linguistics at the University of Arizona and an emeritus professor of linguistics at the Massachusetts Institute of Technology. Ian Roberts is a professor of linguistics at the University of Cambridge. Jeffrey Watumull is a philosopher and the director of artificial intelligence at Oceanit, a science and technology company.

The Times is committed to publishing a diversity of letters to the editor. We'd like to hear what you think about this or any of our articles. Here are some tips. And here's our email: letters@nytimes.com.

Follow The New York Times Opinion section on Facebook, Twitter (@NYTopinion) and Instagram.