Response to Chomsky et al. "The False Promise of ChatGPT"

Wesley Kuhron Jones

Version 1.0.1
This version posted: 2023-11-16
Originally posted: 2023-03-31

Copyright and disclaimer

This document is Copyright ©2023 by Wesley Kuhron Jones.
The document may be freely distributed, provided this copyright notice is included and no money is charged for the document. This document is provided "as is". No warranties are made as to its correctness. The latest version of this document is available at http://wesleykuhronjones.com/ResponseToTheFalsePromiseOfChatGPT.pdf

Changes since version 1.0.0

- Added version number (version 1.0.0 was the document originally posted).
- Added copyright and disclaimer, adapted from "Surreal Numbers – An Introduction" by Claus Tøndering, version 1.7, retrieved 2023-11-16 from https://www.tondering.dk/download/sur.pdf

Copyright ©2023 Wesley Kuhron Jones. All rights reserved.

Here I will argue that ChatGPT and similar models are, at present, as unconscious as a calculator. This is because they have only learned to imitate statistical patterns without any understanding of the world in which those patterns exist. This means that the way they understand and use language (or art, or music, or computer programming languages) is completely different from the way humans do. If AI models are coupled with means of interacting directly with the world physically and attending to their own mental processes, then I would be happy to call them "minds" or "conscious". However, even then, their inner workings will be vastly different from those of humans, and so we should remember that anything they say or do will always be driven by motivations and understandings that are completely alien to us. This applies even for language: even though both humans and AI are catching on to statistical patterns in language, the way that humans do this is motivated by the nature of our shared subjective experience in the world, something that AI will never share with us because of its different implementation in physical substrate. Further, this suggests that we should not trust AI too much in what it says about anything in any domain.

Chomsky et al. (2023) are right that ChatGPT and other machine learning models which seek only statistical knowledge do not function like the human brain. However, Chomsky relies on the idea of the human language faculty working differently from domain-general cognition, which I disagree with. Generativism tries to find out which sentences are and are not grammatical in a language, and to use this to deduce ways of seeing language as a mathematical system. This is not how language works. Grammaticality is gradient, which, granted, plenty of generativists would agree with. But humans use many of the same statistical learning and impressionistic judgments that AI does for producing and evaluating language, rather than deriving these judgments from formal syntax.

It is of course interesting to see how productive a certain structure is, e.g. the "out-" prefix which plays nicely with normal verbs ("I outran her.") but not phrasal verbs ("John out-rang-up Julia." meaning that John did a better job ringing up purchases at a cash register than Julia did). But the feeling of unacceptability of such a sentence is not due to the machinations of abstract syntax in a specialized part of the brain. It is a consequence of a familiarity heuristic (Kahneman 2013), cognitive ease of processing (ibid.), and the reinforcement learning of communicative effectiveness in social situations that creates the culturally-transmitted form-function mapping of language. The familiarity heuristic gives the feeling of "I haven't ever heard someone say something that sounds like that, so I don't like it." Cognitive ease gives the feeling of "That was hard to process, so I don't like it." The

Copyright ©2023 Wesley Kuhron Jones. All rights reserved.

reinforcement reward of communicative effectiveness gives the feeling of "even if I had heard someone say something like that, or if I can think of a structure that it reminds me of, I still am not sure what you mean by saying it that way, so I don't like it and I wish you would say it in a more conventional way."

These impressions can be overridden through enough exposure. This is why bilingualism can lead to structures being calqued between languages with wildly different structures (Ross 2007). Thus, both humans and AI can learn novel patterns that haven't been attested in any language they've ever heard before, and this contributes to many cases of language change. However, Chomsky is right that some patterns are not found in any human language, and AI in principle has no such limitation. This is a reflection of the difference between human brain structure and machine learning algorithms. This is not specific to language; the same could be said of the kinds of patterns that humans like in music, or aesthetics in art, or logic that makes sense to us in mathematical proofs. Here I agree with Chomsky that such things work differently in our brains versus in AI.

ChatGPT, Midjourney, and all similar models know a lot and also know nothing (@willtoulan on Instagram, p.c., 2023). Such a model has memorized large amounts of correlation and nothing else. It has no idea of reference, of what the strings of symbols it uses mean. Nor is it even aware that they mean anything, nor is it aware of the idea of meaning at all. It is a giant calculator.

Imagine you were placed into a purgatory where, in order to get into heaven, you have to correctly predict the contents of a book of numbers that will be given to you in 1,000 years. For now, all that is given to you is the title of that book. To prepare yourself for this task, you have access to a massive library of books of numbers, and as much notebook paper as you want. You spend your time writing down all the patterns you can possibly find to describe what form the books' contents take, and how those relate to their titles. You can never be completely sure of the patterns you find though, only confident to various degrees. There are no rules, like a cellular automaton or a perfect grammar or something like that, that allow you to predict the output with 100% certainty if only you find the correct set of rules. Instead, there are just tendencies of various kinds. If you succeed at constructing the correct contents for the book, then you have learned the correct statistical patterns and can go to heaven. If, once you got to heaven, you were told that the library was actually the collected works of all marine biologists who wrote in Spanish, you would have never

Copyright ©2023 Wesley Kuhron Jones. All rights reserved.

suspected such a thing. As far as you knew, it was all just a bunch of numbers that had some trends in them.

This is how machine learning thinks. If ChatGPT had an internal monologue when it is composing a response, it would be something like this: "All right, for the next letter, I'm really feeling like it should be 'e'. For the next letter, I could see it maybe being 's', but I feel like it's probably 'r'. For the next letter, I'm REALLY feeling like it's 'a'! For the next letter, ...". It has no grasp of anything that the text is about, only impressionistic predictions about which strings of letters feel better than others. It has read various things about chess, but when asked to play a chess game, it failed hilariously. It has no ability to recall the things it has already read about chess and use those to make decisions.

How is this different from the human mind? Is it? Linguistic knowledge is much like the trends that the robot learns, rather than rules that are followed to derive some structure from another. We absorb statistical patterns, and they are reinforced by social acceptability, communicative effectiveness, ease to produce with your mouth, ease to hear, ease to follow with your working memory, and various other factors that relate to our most basic hardware and software as social beings with sensory inputs. For this, we use the same kinds of pattern recognition and association that we use in other domains such as music, art, figuring out who likes who in a social network, and everything else in our experience. So a machine learning model should also be able to deal with all of these, and they already can.

However, while humans do produce similar output to each other because of mutual imitation, we also do this because we share the same underlying motivations and thought processes that drive us to create certain kinds of output in the first place. Two human babies raised together, isolated from all other humans, would create a language with certain features and expressive capabilities because of the kind of brain they have and the kind of world they live in. Machine learning models have only imitation; without it, they have no instincts, no motivations, no reason to produce a certain output over another. Even when their output is remarkably similar to what humans would produce, humans and machines are producing these outputs for fundamentally different reasons.

I do agree with Chomsky that our minds are interested in "[seeking] not to infer brute correlations among data points but [creating] explanations". We want to know what things mean, why they are true, how they relate to the world that we live in, how they are or are not consistent with stories we have told ourselves based on past experiences. Could machine learning ever do this? Without a connection to the actual world in which the things it learns

Copyright ©2023 Wesley Kuhron Jones. All rights reserved.

about exist, I don't think so. It is the difference between reading as much as you can about Spain, and actually going to Spain. There is an incredible difference: the latter feels very rich and allows you to "feel" the facts that you have learned in a new way that you never would have gotten if Spain hadn't become part of your direct sensory input (Heidegger). Things that you knew declaratively from study suddenly make sense, you have a sense of *why* they're true now, or you can *feel* that they're true, whereas before you only had memorized *that* they were true. This same mechanism is behind the efficacy of visualization as a tool for developing intuition about mathematics, for example.

I don't think that our minds' concepts of explanation, or the other such things that I've listed, are out of the reach of machine learning. In fact, our brains are just computers. I don't believe in strong forms of substrate independence (Bostrom 2001), because the hardware we have is so different from electronic circuitry. It in fact relies on things like neurotransmitters leaking around by random fluid motion, whereas a computer is built to have maximum predictability in the behavior of its hardware. Such differences affect what kinds of programs each substrate can reliably run. Aside from that, though, our brains are still computers. They are physical systems from which all of our mental experiences emerge as an illusion. Our consciousness is an artifact of the attention mechanisms that allow us to focus on certain things, think "about" something, work on tasks, etc. (Graziano 2018) Similar attention mechanisms are precisely what modern Transformer neural networks have leveraged to become so successful (Vaswani et al. 2017). There is no objective sense in which to define something as conscious or not; this too is gradient, and it takes many flavors, a vast space of which the set of human experiences is only a small part.

If "having an internal experience" arises from having attention to one's own cognition, then it is a small jump from the Transformer to this. To my knowledge, Transformers are only attending to parts of the input, not to their own cognition. Even attention to a hidden layer does not seem to me to count as metacognition, because it is still only looking at partially-digested trends in inputs, rather than at its own process for doing that digesting. But it is pretty easy to conceive of the addition of metacognition or executive function like this to their operation. At that point, I'd say they have an internal experience, albeit one incredibly alien to us, like the Solarian Ocean (Lem 1961) whose motions will forever remain inscrutable. Once they gain sensory input and the ability to directly interact with the world rather than data that is handed to them as a secondary source, they will have joined us in having a mind (Hofstadter 2007).

Copyright ©2023 Wesley Kuhron Jones. All rights reserved.

That said, will they understand language? Will they understand the world as we do? I don't think so. The algorithm they are running is so different from ours that, while they would develop *some* understanding of the world given the right (meta-)cognitive mechanisms and sensorimotor tools to interact with things in the physical realm, their internal representation of things will be very unlike ours. We have all experienced this sort of seemingly unbridgeable mismatch in cognitive schema about something. It arises from different representations of data and different algorithms used when handling it. This effect of talking past each other is seen with a relative who is on the opposite side of the political spectrum, a person from a different culture when discussing religion, a cat who thinks your pointing gesture means to look at your finger rather than what you are pointing at, a dog who thinks you doing yoga is some kind of game, a linguist who thinks syntax has a deep structure that transforms into a surface structure, or any other being whose assumptions and means of operation are different from yours to an extent that prevents mutual understanding ("If a lion could talk, we could not understand him" (Wittgenstein 1953)). The machine learning algorithms, once they have something that could be called "understanding" as I have described above, will be like this but to a much greater extent. So we should treat what they say as such, coming from a mind who sees the world vastly differently, and we should be aware of the implications of that for how much stock we put into their ideas.

Acknowledgments

Thanks to Yiding Hao, Trevor Norris, and Vsevolod Kapatsinski [order chronological] for their comments on earlier drafts.

Trevor Norris's recommendations for further reading:

For an angle you’re probably not looking at too much, but with some similarity, I would recommend Reality is Not What it Seems by the physicist Carlo Rovelli and the works of the philosopher Jay Garfield on Madhyamaka Buddhism (as well as his translations of the writings of Nagarjuna). I think these are somewhat in line with the whole Gödel-Escher-Bach thing about distinctions between categories being actually rather fluid and not as rigid as one may ordinarily think.

Copyright ©2023 Wesley Kuhron Jones. All rights reserved.

References and intellectual inspirations

Disclaimer: I have not read most of these works, only pieces of them or things written about them by others. I use three asterisks to mark works that I have read fully, two to mark works that I have only partially read, and one to mark ones that I have not read in the original at all.

** Bostrom, Nick. 2001. Are You Living in a Computer Simulation?

*** Chomsky, Noam, Ian Roberts & Jeffrey Watumull. 2023. The False Promise of ChatGPT. *The New York Times*.

** Graziano, Michael S. 2018. The Attention Schema Theory of Consciousness. In Rocco J. Gennaro (ed.), *The Routledge Handbook Of Consciousness*, 174–187.

* Heidegger's phenomenological approach emphasizing that consciousness is epiphenomenal to the human's existence together with the world around us.

* Hofstadter, Douglas R. 2007. *I Am a Strange Loop*.

*** Kahneman, Daniel. 2013. *Thinking, fast and slow*. 1st pbk. ed. New York: Farrar, Straus and Giroux.

* Lem, Stanisław. 1961. *Solaris*.

*** Ross, Malcolm. 2007. Calquing and Metatypy. *Journal of Language Contact* 1(1). 116–143.

** Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser & Illia Polosukhin. 2017. Attention Is All You Need.

** Wittgenstein, Ludwig. 1953. *Philosophical investigations*.

Copyright ©2023 Wesley Kuhron Jones. All rights reserved.