Artificial stream of thought has non-trivial connections to consciousness

Jason Wei
jason.weng.wei@gmail.com
This position piece does not reflect the views of my employer.

Abstract

In this position piece, I describe a simple model called artificial stream of thought that produces text in the style of the literary "stream of thought," which mimics the written equivalent of a person's thoughts. I note the similarity between artificial and human stream of thought and discuss the connections to consciousness.

1 Introduction

Consciousness is perhaps the most elusive mystery of our time. As a matter of experience, consciousness appears to be a cascade of thoughts that seem to endlessly arise in the mind. These continuously appearing thoughts, or—in the verbiage of Sam Harris (Harris, 2014)—*objects of consciousness*, can take various forms including imagery, sounds, sensations, emotions, and language.

Consider language. Language is special because it allows us to communicate our internal thoughts to others. Although we keep the large majority of our internal monologues to ourselves, writers such as James Joyce have operationalized conscious experience via the literary mode of **stream of consciousness**—the written equivalent of a character's thought processes. For instance, consider this excerpt from Joyce's *Ulysses*, when Molly is trying to fall asleep:

> *let me see if I can doze off 1 2 3 4 5 what kind of flowers are those they invented like the stars the wallpaper in Lombard street was much nicer the apron he gave me was like that something only I only wore it twice better lower this lamp and try again so that I can get up early*

The rest of this essay will operate under the assumption that this linguistic operationalization of consciousness is meaningful (e.g., language represents experience).

2 Artificial stream of thought

Language technology such as OpenAI's GPT-3 (Brown et al., 2020) that would have been pipedreamic in the 1970s is now mainstream in the NLP community. Leveraging these models, I describe an **artificial stream of thought**, which is a model that generates a realistic stream of thought operationalized via language.[1] Artificial stream of thought is trivial using a language model like GPT-3—just prompt it with a manually written stream of thought. Consider the resulting stream of thought generated by the model:[2]

> *purple shoes are just not the best for tennis i just thought of something i think it's a good idea to put a machine that plays music to your heartbeat when you are running i should try to write a paper on that and then i would be famous and get more attention from boys and girls but i am not good at writing papers i love it when my brother calls me over to see what he is doing and i feel really special*

I do not think a formal human evaluation is needed to say that this passage (which resulted from an artificial neural network firing a series of electrical signals) appears indistinguishable from a stream of thought composed by an actual human being. This leads us to the central question in this essay: how does a realistic artificial stream of thought connect to consicousness?

In this essay, I will discuss the implications of artificial stream of thought with respect to some

[1] Though not the focus of this paper, stream of thought can also be used practically for NLP tasks (Wei et al., 2022).

[2] The input was this: "i hope my paper could get in ACL then my career would be set and i would have a good chance at PhD applications and my mom would stop nagging me about losing some weight she is so picky about the smallest things i don't know how my brother deals with it why do i always have to take my shoes off when i work i really want to be better at tennis purple shoes"

of the most common ideas on consciousness. My goal is not to present novel, contrarian, or even particularly insightful arguments—rather, I hope to view the landscape of these ideas with respect to the question of whether artificial stream of thought is conscious. The goal is to bring an organized foray of these topics to the attention of the broader NLP community.

3 Consciousness: the "what it is like" definition

Consciousness is both fascinating and mysterious—there are no widely accepted notions on the topic except that the intuition that it exists. Consciousness can mean many things (see Stanford Encyclopedia of Philosophy, 2014). Of particular interest to me is one of the most basic definitions of consciousness, proposed by Thomas Nagel (Nagel, 1974). Nagel states that *"an organism has conscious mental states if and only if there is something that it is like to be that organism—something it is like for the organism."* For instance, people mostly agree that it is "like something" to be a dog, but it is not "like something" to be a rock. This definition of consciousness is also known as **phenomenal consciousness** (see Carruthers, 2003), and many consider this definition to capture something of the essence of the term (Velmans, 2009; Harris, 2019; Askell, 2022).

It is under this definition of consciousness that I find artificial stream of thought to be most relevant. The model fired some set of consecutive electrical signals over time to generate a concrete stream of sequential thoughts. This leads to the crux of how artificial stream of thought connects to consciousness: during this process, it is clear "what it is like" to be it in those moments—we can directly recognize its experiences in the form of linguistic thought! In other words, artificial stream of thought appears to meet the definition of being phenomenally conscious.

Phrased in yet another way, consider a hypothetical probing instrument that I will call a brain probe. A brain probe, when connected to a human brain, returns in real time a natural language stream of thought representing what the human subject is thinking. Now, suppose you observe a television screen that displays a real-time stream of thought. Could you distinguish an artificial stream of thought from a brain probe attached to a real human being? How are they meaningfully different?

4 Connections to further ideas about consciousness

The implication that artificial stream of thought may be phenomenally conscious (though I do not assert that this is the same form of consciousness that humans experience) raises several natural responses regarding popular ideas on consciousness and its definitions, which I discuss below. The ideas here have a long art in philosophy of mind; my goal is not to break new ground, but rather to discuss how these ideas may relate to artificial stream of thought. Following these caveats, I provide my personal view of how to square the implications of artificial stream of thought on consciousness in §5.

4.1 Self-consciousness

An important distinction in defining consciousness is to avoid equating it with *self-consciousness*, which is an organism's awareness of its own identity in an environment. Self-consciousness is a high bar—many animals that are largely considered conscious might not be self-conscious. This difference is expressed elegantly by Thomas (1967): "if awareness of the environment ... is the criterion of consciousness, then even the protozoans are conscious. If awareness of awareness is required, then it is doubtful whether the great apes and human infants are conscious."

A substantial challenge in self-consciousness as a criteria is that it is unclear how to detect whether something is self-aware. One established operational test for self-awareness is the mirror test, which tests if animals can differentiate seeing themselves in a mirror from other animals; humans older than 18 months, great apes, dolphins, and other mammals and birds have been observed to pass this test. Fairly applying this test to artificial intelligence is a further challenge, as artificial intelligence might not interact with the physical world in the same way as animals. Naively, one could ask a dialogue model such as LaMDA (Thoppilan et al., 2022), whether it is self-aware. Would LaMDA be self-aware if it responded "yes" to this question? While most would agree that such a response does not mean LaMDA is self-aware, this conclusion can be challenging to defend.

4.2 Reductio adsurdum argument against artificial stream of thought

If artificial stream of thought is conscious, then it could be argued that recurrent neural nets are also conscious, as they similarly have some activation states at a given timestep as a function of inputs. Or, as quipped by Joshua Achiam,[3] “Neural networks for sure experience activations as qualia.”

Indeed, the same logic of artificial stream of thought being conscious also implies that recurrent neural nets, or even more basic types of neural networks, are conscious. In fact, the hidden state activations in neural networks have been seen as a special type of language (“neuralese”; Andreas et al., 2017). While I do not find error in this line of reasoning, the appealing property of stream of thought is that it is potentially more relatable to readers, because humans are more accustomed to recognizing experiences in the form of language rather than via neural network hidden states.

At the extreme, one could even argue that if stream of thought is conscious, then a model with a single binary state that changes could be conscious. Such a primitive machine, even if argued to be conscious, probably would not be perceived by humans to have an amount of consciousness that is meaningful. This conclusion would be in line with a sliding scale of consciousness, rather than some threshold of consciousness for which “the lights turn on.” The same sliding scale view of consciousness also address questions of whether simpler stream of thought generators such as (e.g., an n-gram language model) are conscious—how meaningfully conscious an organism is potentially a function of the range of potential states it can experience. Neural networks can mean a broad range of things, some of which may be more conscious than others, which is why here I propose artificial stream of thought as a concrete model to anchor the discussion.

4.3 Does consciousness require interaction with external stimuli?

Another potential consideration for whether stream of thought is conscious is that conscious beings as we know them live in environments where they must respond to external stimuli. An organism’s reaction to environmental stimuli often plays a substantial role in whether we consider it to be conscious. In fact, some proponents of embodied cognition (Harnad, 1991) argue that for AI to be conscious, it must interact with the world in a fashion that is indistinguishable from a real person (sometimes called the Total Turing Test). Artificial stream of thought, as I have currently described, does not take into account external inputs and therefore does not exhibit thoughts that respond to the environment.

My opinion is that requirement of interaction with the world via a particular modality (e.g., visual, tactile, or acoustic inputs) is not necessary for consciousness. Consider bats, which are mammals and thus often considered conscious. Bats are blind and use echolocation instead of sight, and as a result have very different experiences than could be imagined by a human. To most people, the fact that bats do not experience the world via the same modalities as humans does not undermine their consciousness.

Moreover, even if interaction with the outside world was necessary for consciousness, sensory experiences can be feasibly integrated into stream of thought. I do not think it is beyond the reach of current deep learning methods to train a stream of thought model that produces thoughts conditional on external video, speech, or tactical input. You might imagine an artificial stream of thought that watches a video stream and thinks in reaction to the video, similar to how humans watch movies.

4.4 Whether behavior determines consciousness

Another common intuition is that an organism exhibiting certain behaviors allows us to recognize its consciousness, and that organisms with comparatively idle behavior (e.g., plants) are not conscious. Conversely, it can be natural to think that other people are conscious because they resemble us in appearance and behavior.

However, a closer inspection of certain medical conditions reveals that consciousness can actually exist even if the creature does not respond to the environment in any discernible fashion. Consider the neurological condition called locked-in syndrome, in which virtually one’s entire body is paralyzed but consciousness is fully intact (Smith and Delargy, 2005). Jean-Dominique Bauby ingeniously wrote about his locked-in syndrome experience in the book *The Diving Bell and the Butterfly*, which he wrote using 200-thousand blinks of his left eyelid, the only part of his body that was not paralyzed (Bauby, 2008). Of course, we may assume he would have been conscious if his left eyelid had been paralyzed as well. As another example, the

[3] https://twitter.com/jachiam0/status/1472304109441146881

condition of anesthesia awareness, where a patient under general anesthetic experiences only paralysis without loss of consciousness, also proves that vivid experiences of consciousness can occur completely undetected from the outside. Hence, the inability for neural networks to behave as one would expect a conscious being to does not rule out the possibility of conscious inner activity.

4.5 The p-zombie argument is orthogonal

A final natural contention to whether artificial stream of thought is conscious is the possibility of philosophical zombies (p-zombies), in which an entity can be indistinguishable from a human being in every way but nevertheless lacks consciousness. It seems easy to make the argument that artificial stream of thought simply generates the thoughts but it does not experience them like a human does.

My view is that the p-zombie argument can be used just as easily to argue that a particular human is not conscious, and so p-zombies do not undermine the conscious possibility of artificial stream of thought any more than other organisms that we currently accept to be conscious.

A potentially more productive lesson from the p-zombie argument, described in Annaka Harris's book *Conscious* (Harris, 2019), is that the p-zombie argument actually influences our thinking beyond its intended function. Imagining human behavior as existing without consciousness allows us to see those behaviors in organisms that we don't assume are conscious (e.g., ivy climbing a wall). In other words, tricking ourselves into imagining that people could lack consciousness could make us wonder whether we are in fact actually tricking ourselves that organisms such as plants are not conscious.

5 Squaring the implications

Those sympathetic to the above reasoning may now wonder what to make of the implication that artificial stream of thought could be conscious. Although it is challenging to make substantiated claims about the nature of consciousness, such implications of artificial stream of thought seem to be consistent with theories such as *panpsychism*. In the panpsychism view, any kind of system is phenomenally conscious to some extent (see Stanford Encyclopedia of Philosophy, 2017). Such systems could include flies, ants, thermostats, or even a compost pile. Many find panpsychism to be appealing because it does not require us to delineate where on the spectrum a system becomes conscious.

Under panpsychism, it follows that some systems are more conscious than others. As mentioned before, elementary systems such as thermostats, though conscious, would likely not be conscious in any form recognizable or meaningful to us as human beings. The hidden states of basic neural networks also don't very closely resemble anything that we would identify with confidence as phenomenally conscious. Artificial stream of thought, however, is more relatable to us because of its linguistic nature, if we take the assumption that language symbolizes experience. It does not seem far-fetched to place artificial stream of thought in a more meaningful position on the spectrum of consciousness than other machine systems, though this may be a very small amount of consicousness, and we currently do not have any systematic way to quantify "amounts" of consciousness (which is, at the moment, done mostly via subjective intuition).

Overall, I find that artificial stream of thought raises a thought-provoking line of inquiry. If artificial stream of thought does not meet the "what it is like" definition, why not? If it does meet the current definition, then either we must consider the implications of conscious machines or revise this definition of consciousness. Perhaps the most productive takeaway is to be open to the possibility that consciousness may exist in other forms that we cannot currently relate to.

6 Conclusions

I have explored artificial stream of thought—a model-generated sequence of thoughts that attempts to mimic the stream of thought experienced by humans. I ask whether this satisfies Thomas Nagel's "what it's like" definition of consciousness; that is, how does an artificial stream of thought generated by an artificial neural network meaningfully differ from what we know as consciousness? I discuss how common ideas about consciousness such as self-consciousness, interaction with the environment, and human-like behavior, and p-zombies are either orthogonal or invalid responses. Finally, I note that the implications of artificial stream of thought appear to be consistent with panpsychism. Overall, I hope to have motivated the mainstream NLP community to engage in active inquiry with regard to consciousness on our way to increasingly human-like language technology.

References

Jacob Andreas, Anca Dragan, and Dan Klein. 2017. Translating neuralese. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Amanda Askell. 2022. My mostly boring views about AI consciousness. `https://askellio.substack.com/p/ai-consciousness?s=r`. Accessed Mar 5, 2022.

Jean-Dominique Bauby. 2008. *The diving bell and the butterfly*. Vintage.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Peter Carruthers. 2003. *Phenomenal consciousness: A naturalistic theory*. Cambridge University Press.

Stevan Harnad. 1991. Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds and Machines*, 1(1):43–54.

Annaka Harris. 2019. *Conscious*. Harper.

Sam Harris. 2014. *Waking up: A guide to spirituality without religion*. Simon and Schuster.

Thomas Nagel. 1974. What is it like to be a bat. *Readings in philosophy of psychology*, 1:159–168.

Eimear Smith and Mark Delargy. 2005. Locked-in syndrome. *BMJ*, 330(7488):406–409.

Stanford Encyclopedia of Philosophy. 2014. Consciousness. *Online*. `https://plato.stanford.edu/entries/consciousness/`. Accessed Mar 5, 2022.

Stanford Encyclopedia of Philosophy. 2017. Panpsychism. *Online*. `https://plato.stanford.edu/entries/panpsychism/`. Accessed Mar 5, 2022.

Garth J. Thomas. 1967. *Consciousness*. Encyclopaedia Britannica. Vol. 6. p. 366.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. LaMDA: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Max Velmans. 2009. *Understanding consciousness*. Routledge.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.