

Simulating Understanding with Deep Learning

By Catherine Li

AUTHOR BIO

Catherine Li is an aspiring neuroscientist and student at Newport High School in the US. Beginning from an early interest in the human mind, she took classes on biology, neurology, psychology, and recently interned at a research lab, Scintillon Institute, in San Diego where she gained hands-on experience in benchwork. Her interest in the mind extends to character analysis in books, movies, and TV shows which she enjoys both inside and outside of English class.

ABSTRACT

Artificial sentience is a significant ethical topic that would require talks of AI/AS rights and change how the world has begun to use AI in daily life. With developments in neural networks and the new release of ChatGPT, that topic is more relevant, but not necessarily a real concern yet. Current signs used to determine animal sentience and awareness cannot be applied to AI the same way, but the algorithms and capabilities of current AI can be examined instead. When tested on basic logic problems, ChatGPT and other large language models fall short. Current AI is not capable of true logic or understanding that comes from consciousness, revealing that the simulation of understanding has not become reality.

Keywords: AI, Sentience, Deep Learning, ChatGPT, Consciousness.

INTRODUCTION

Many people have wondered about the possibility of conscious and/or sentient AI. There are countless sci-fi depictions of them, from both a sympathetic view and a fearful view, but they're just that: fiction. AI has advanced significantly, with complex neural networks that mimic the neurobiology of organisms. Last year, ChatGPT was released and became famous over its capability to understand and answer prompts, leading many people again to question the minds of AI.

Deep Learning

With more complexity than machine learning, deep learning (DL) is made of neural networks that have many layers. The 1st layer is an input layer that takes the data as numbers. Each number is an input "neuron" that's connected to varying amounts of other neurons in the 2nd layer that each multiply the input number by the weight of the connection. Within the 2nd layer neuron, the now weighted input is added to the bias and the result is often put in a ReLU, sigmoid, or tanh function to produce the output. The output is then passed as an input to the following layer of neurons with their own weights and biases. All deep learning algorithms have many connected hidden layers that have tuned their weights and biases through stochastic gradient descent, eventually leading to the final output layer that gives the result. Deep learning is already capable of many things, like image processing, understanding speech, diagnosing medical conditions, etc.

Consciousness and sentience

Consciousness is defined as the state of being aware of internal and external existence, while sentience is defined as the ability to experience emotions and sensations, and all sentient beings have consciousness. When determining animal consciousness and

awareness there are different benchmarks used, like mirror self-recognition, but most focus on animals' capacity to suffer, signs of pain or distress, or its neuroanatomical similarities to the human brain (Proctor, 2012). Despite the inspirations from neurology, these aren't as applicable to an AI, which doesn't have a brain and can be directly programmed to say something or not say something. But with the complexity of modern AI and neural networks that are capable of passing the Turing test (Jannai, 2023), people begin to consider the possibility that there is awareness in these algorithms and these simulations of emotion and understanding could be real.

The Chinese Room Argument

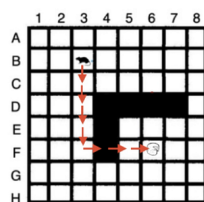
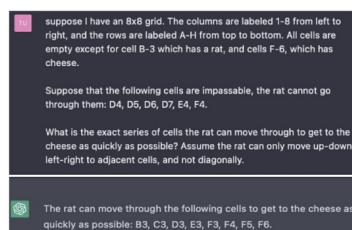
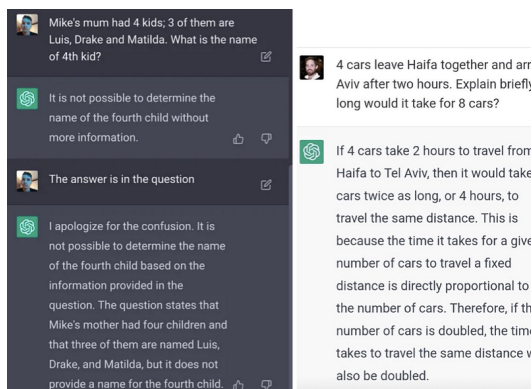
In the end, a chatbot doesn't understand the meaning of the words it's telling you. Instead, it knows that a certain word will typically be said after another word which will typically be said after another word and so on until it forms a response that would likely be said after your input. If you asked it the meaning of something, it could answer using the same probabilities and linking together. Following the Chinese Room Argument made by John Searle (1980, 1990), a person who doesn't understand a word of Chinese could be given a set of syntactical rules for Chinese to the point of holding a grammatically and logically correct Chinese conversation with an actual Chinese speaker. But the person still isn't a Chinese speaker because they don't understand the conversation, they're simply following the rules without knowing the meaning of any of the words. When seeing 苹果 for example, the person in the room will not actually know that means apple and understand. But with the right instructions on syntax, they could produce a response in Chinese that makes it seem like they do. AI functions this way, chatbots know syntax rules and what's supposed to go together, but they don't know why.

Natural Language Processing, Understanding, and Generating

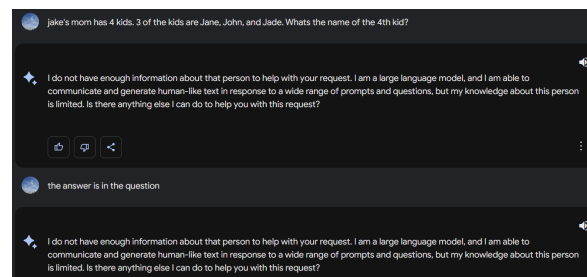
ChatGPT will give a perfect response when asked the meaning of a word, but it can do that because it's found a pattern to how other people respond when asked the same question and uses that pattern. Simply put, Natural Language Processing works by tokenizing every word into a matrix (vector) that's stored in the "neurons" which each hold 1 value of a matrix. Relationships between words are represented in the relationships between the vectors, such as king – man + woman = queen.

These matrices are altered throughout the hidden layers and analyze the test with Natural Language Understanding, which detects patterns and words that are associated with other specific words to determine things like sentiment analysis. And in Natural Language Generation, trained with large datasets of human writing, the AI will find the right patterns and words that form the typical responses people say, not necessarily the correct ones.

Returning to ChatGPT, there are various examples of logical errors it has made.



In all these instances, a person could imagine the scenario given and figure out the answer using an actual understanding of what's being said. Meanwhile, ChatGPT doesn't have that understanding despite the name "natural language understanding." In the first example, a person can imagine the given scenario and understand that there is a family with four kids named Mike, Luis, Drake, and Matilda. ChatGPT's answer shows it lacks that understanding despite the simplicity of the question. It can do powerful syntax manipulation with pattern recognition, but doesn't have a mind to do true thinking and understanding. This isn't limited to ChatGPT. Bard, for example, has similar issues.



Conclusion

Current AI is certainly not conscious, sentient, or aware, but the neuroscience behind consciousness and sentience is still unknown. It's impossible to know what exactly another person/animal/being is experiencing, which is why thought experiments about the possibility of one person's red being another person's blue exist. So when it comes to the consciousness of AI, it's hard to know what's really possible. Since future AI may overcome the issues of current AI with other neurology inspired algorithms, other developments are still an ethical concern.

REFERENCES

Cosmides, L., & Tooby, J. (2000). Evolutionary psychology and the emotions. In M. Lewis & J. M.

Haviland-Jones (Eds.), *Handbook of emotions* (2nd ed., pp. 91–115). New York, NY: Guilford Press.

Tooby, J., & Cosmides, L. (2008). Evolutionary Psychology and the Emotions and Their Relationship to Internal Regulatory Variables.

https://www.researchgate.net/publication/281273968_Evolutionary_Psychology_and_the_Emotions_and_Their_Relationship_to_Internal_Regulatory_Variables

Silva, B. A., Gross, C. T., & Gräff, J. (2016). The neural circuits of innate fear: detection, integration, action, and memorization. *Learning & memory (Cold Spring Harbor, N.Y.)*, 23(10), 544–555.

<https://doi.org/10.1101/lm.042812.116>

Lavelle S. (2020). The Machine with a Human Face: From Artificial Intelligence to Artificial Sentience.

Advanced Information Systems Engineering Workshops: CAiSE 2020 International Workshops, Grenoble, France, June 8–12, 2020, Proceedings, 382, 63–75.

https://doi.org/10.1007/978-3-030-49165-9_6

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424. doi:10.1017/S0140525X00005756

Walter, Yoshija & Zbinden, Lukas. (2022). The problem with AI consciousness: A neurogenetic case against synthetic sentience.

10.48550/arXiv.2301.05397.

Searle, John (January 1990), "Is the Brain's Mind a Computer Program?", *Scientific American*, vol. 262, no. 1, pp. 26–31, [Bibcode:1990SciAm.262a..26S](#), [doi:10.1038/scientificamerican0190-26](#), [PMID 2294583](#)

Allen, C. & Hospedales, T. (2019). Analogies Explained: Towards Understanding Word Embeddings. *Proceedings of the 36th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research* 97:223–231 Available from <https://proceedings.mlr.press/v97/allen19a.html>.

Jannai, D., Meron, A., Lenz, B., Levine, Y., & Shoham, Y. (2023, May 31). *Human or not? A gamified approach to the turing test*. arXiv.org. <https://arxiv.org/abs/2305.20010/>.

Proctor, H. (2012). Animal Sentience: Where Are We and Where Are We Heading? *Animals*, 2(4),

628–639. MDPI AG. Retrieved from

<http://dx.doi.org/10.3390/ani2040628>

Melcher, K. (2021). Example of a fully connected, feedforward neural network [Image]. Retrieved from <https://www.knime.com/blog/a-friendly-introduction-to-deep-neural-networks>.

(2019). Diagram of vectors demonstrating how relationships between words can be shown [Image]. Retrieved from

<https://www.ed.ac.uk/informatics/news-events/stories/2019/king-man-woman-queen-the-hidden-algebraic-struct>.