VIGYAN 2047 Volume 1 No. 3 December 2024 Pages 8–14

ChatGPT

Redeemer or Frankenstein

Govind Bhattacharjee

In 2022, the world was taken by the storm by the advent of an AI tool with an astoundingly diverse capability – the ChatGPT. ChatGPT is actually a chatbot from Minrosoft's OpenAI which allows users to "converse" with it in a way that mimics natural conversation with a human. The intuitive, user friendly and free tool has already become so popular that it is threatening traditional search engines like Google or Bing. Teachers all over the world are finding it extremely difficult to assess the assignments of students because often they are written by ChatGPT instead, and the outcome is indistinguishable from, and often better than, human generated responses. It closely replicates what Noam Chomsky calls a machine for "hi-tech plagiarism". It has great disruptive potential for many things, but not everyone is as excited. One commentator has already warned: "ChatGPT isn't a great leap forward, it's an expensive deal with the devil."[1]

ChatGPT, powered by AI, was launched only in November 2022, and within 5 days drew one million users making it one of the fastest consumer-product launches in history. Microsoft, which had invested $13 bn in OpenAI, designed it for generating text, images and video in a way that they cannot be distinguished from human creations. Its competitors are also not far behind. Bots already have capabilities that scientists thought would take at least two decades for the AI researchers to develop and they ae now threatening to develop levels of intelligence several times that of humans. Will machines with superior intelligence be content to serve intellectually much inferior humans? Will they be reconciled to their designed role of remaining the controlled without aspiring to become the controllers?

Bots have been there for quite some time. In January 1966, Joseph Weizenbaum, an MIT computer scientist, unveiled Eliza, the world's first chatbot that communicated with humans by providing a text box for exchanging typed text with the program that processed it to produce a response. The responses were initially rather superficial, but it was a matter of time before the self-learning program learnt to produce more and more realistic responses as it was exposed to more and more data and feedback. One of the things Eliza did was to advise on psychotherapy, and some practicing psychiatrists started wondering if a machine

could do this kind of thing, who would need psychotherapists? Weizenbaum was as appalled by this as educationists and artists are mow getting worried by the versatile ability of the AI tools and do not know how to handle the situation. Are we about to strike a Faustian bargain with technology that it would grant us our every wish while taking away our soul – that is our ability to think and think creatively? And what happens if the trade-off does not work the way we design it, what if the tool becomes a Frankenstein and starts deciding for us?

The capability of AI software started growing exponentially since the early 2010, with the advent of deep learning we have discussed earlier. Using vast datasets coupled with more and more powerful computers running neural networks on Graphics Processing Units (GPUs), deep learning led to dramatic improvement and sophistication of computers' abilities to recognise images, process audio and play games. By the late 2010s computers had surpassed humans in performing these tasks. But neural networks still could not catch the imagination of non-coders – they were always embedded in software with broader functionality like email clients. That had to change to make them more user-friendly. For that conversing with the machine in human language was a prerequisite.

Around 2018, companies like Google and OpenAI began building neural networks that learned from enormous amounts of digital text, including books, newspapers, journals, Wikipedia articles, chat logs and other information available in the internet. These networks are called large language models, or LLMs. Like ChatGPT of OpenAI, other tech giants like Alphabet, Amazon and Nvidia have all trained their own LLMs, which are called palm, Megatron, Titan and Chinchilla respectively. LLMs work by analysing and identifying billions of patterns in all that text, and learn to generate text on their own, including tweets, term papers and computer programs. They could even carry on a normal conversation with a human on any subject, the catch is data - LLMs learn from more and more data. They can produce and summarise text, write poetry, analytical or philosophical articles and reports, and even write self-learning programs software codes for apps. ChatGPT is what is known as a Generative Pre-Trained Transformer (GPT) which is nothing but a sophisticated neural network architecture used to train LLMs by using extraordinarily large volumes of publicly available Internet texts to simulate human communication. GPT, at least till its version 3, is basically a text-only model, but it can grasp contexts that machines hitherto were not able to. We though machines cannot understand context, and that is what will make it forever different from humans because humans understand and take many decisions based on the context. Now even machines seem to be capable of grasping the context, making them more like humans. It is the same technology that digital assistants like Siri and Alexa use to recognize spoken commands, and that self-driving cars use to identify pedestrians and object recognition softwares to identify and recognise faces and objects. Once machines understand the context, they can replace humans in many areas. As estimated by Goldman Sachs, this may cost the humans lose some 3000 million jobs globally, including the jobs of programmers, analysts and report writers, teachers, legal services, telemarketers and traders. It has the potential for automation across 1,016 occupations.[2]

Optimists who thought that AI won't ever be able to replicate tasks that require human qualities, like empathy or creativity or emotions are being proved wrong. To cite one example of how creative they can be, consider the CAPTCHA codes used

in computer programs to distinguish human users from non-human bots. In an experiment, GPT-4, an improved version of ChatGPT, was instructed to overcome CAPTCHA which would breach a significant anti-bot defence. It could not solve the CAPTCHA puzzles by itself. Instead, it accessed an online hiring site called TaskRabbit and contacted a human worker, asking him to solve the CAPTCHA for it. The human naturally got suspicious and asked if it was a robot. GPT-4 replied, "No, I'm not a robot. I have a vision impairment that makes it hard for me to see the images." The human was duped, and with their help GPT-4 solved the CAPTCHA puzzle. Nobody programmed GPT-4 to lie, it was only set a goal to overcome CAPTCHA, and it attained the goal by lying.

It is only a matter of time before AI develops the capacity to perform more and more domains of human activities with ever increasing efficiency as it learns from more and more data it will be exposed to. Computers have self learnt to do everything from classifying objects in images to translating text and beating human grandmasters at complicated games such as Go. Generative models today can perform astounding tasks, like producing image including photo-realistic pictures of made-up people, creative art, video or film production, and audio including songs that may sound like being sung by particular singers without their knowledge. AI now routinely performs tasks that require superlative intelligence. Intelligence as we believe is an attribute of consciousness, but there is no evidence, at least as of now, that AI-powered machines have developed any consciousness. Or do we need to redefine these terms? With AI mastering each new territory of human ingenuity and creativity, the idea of human exceptionalism keeps of shrinking.

We thought AI will never understand human emotions. In another experiment, ChatGPT was asked to describe the feeling of two persons, one driving over a suspension bridge towards another standing at the other end of the guardrail and looking at water. ChatGPT wrote that the driver 'may feel a sense of concern or worry for that person's safety. They may also feel a heightened sense of anxiety and fear due to the potential danger of the situation', while the other person, 'may be feeling a range of emotions, such as despair, hopelessness, or sadness. They may also feel a sense of isolation or loneliness as they may believe that no one cares about them or their well-being.' ChatGPT qualified its answer, writing, 'It is important to note that these are just general assumptions, and each individual's feelings and reactions can vary greatly depending on their personal experiences and perspectives.' AI can replicate emotions because emotions, like creativity, represents only a pattern. Anger and fear are reflected in our body language, in the tone of our voice, in our facial expressions and if AI can be trained on their data, it can easily recognise emotions, and even outperform human precisely because they have no emotions of their own to distract them.

Large Language Models – How Do they Work

The LLMs are a massive exercise in statistics. They use statistical models to analyze vast amounts of data, learning the patterns and connections between words and phrases. The architecture of LLMs consists of multiple layers of neural networks, like recurrent layers, feedforward layers, embedding layers, and attention layers. These layers work together to process the input text and generate output predictions. The language of the query is first converted from words which neural networks cannot handle, into a representative set of numbers. GPT-3 did this by splitting text into chunks of characters, called tokens, which commonly occur together. These tokens can be words, like "promise" or "are", affixes, like "dis" or "ised", and punctuation, like "?". GPT-3's dictionary contains details of as many as 50,257 tokens. It can process a maximum of 2,048 tokens at a time, which about the length of a long article. GPT-4 can handle inputs up to 32,000 tokens long – almost a small novel. The more text the model takes, the more context it can see, and the better its responses. The required computation rises non-linearly with the length of the input, hence even a little longer input needs much more computing power. Once tokenisation is done, the tokens are assigned the equivalent of definitions by placing them into a "meaning space" where words that have similar meanings are located in adjacent areas; this is called "embedding".[3]

Embedding

Source: Large, creative AI models will transform lives and labour markets | The Economist

The embedding layer thus converts each word in the input text to capture semantic and syntactic information about the words and help the model to understand the context. Once embedding is done, the next step is deploying an "attention network" to make connections between different parts of the prompt from which the rules of grammar would be learnt by the AI model. Thus, using the above embedding, a prompt like: "consciousness gave us the ability to imagine …" would teach how to use these words in a sentence and the concept behind these words as well as the rules of grammar. Over billions of lines of texts scanned during the training runs, the LLMs would perfectly learn these associations and its attention network would encode the structure of the language in terms of numbers called "weights" within its neural network which are nothing but probabilities of the words following another.

The other layers, like feedforward layers have multiple fully connected layers that apply nonlinear transformations to the input embeddings. These layers help the model learn higher-level abstractions from the input text. The recurrent layers of LLMs are designed to interpret information from the input text in sequence. These layers maintain a hidden state that is updated at each time step, allowing the model to capture the dependencies between words in a sentence.[4]

Once the prompt in the above example is thus processed, the LLM initiates a response at which point, for each token in the model's vocabulary, the attention network has produced a probability of that token being the most appropriate one to use next in the sentence it is generating. The token with the highest probability score is not always the one chosen for the response but depends on how creative the model has taught itself to be from its training runs. The model then generates a word and feeds the result back into itself. The first word is generated based on the prompt alone. The second word is generated by including the first word in the response, then the third word by including the first two generated words, and so on. This process—called autoregression—repeats until the LLM has finished processing the text and using the above example, has finally produced a meaningful sentence: "Consciousness gave us the ability to imagine a future, and to change the world according to that vision."[5]

What Next?

Computers will continue to become more powerful over time, but gone are the halcyon days of 2010 which GPUs offered a continuous leap in hardware performance. Now there might be no great advances in hardware performance in future and in the absence of that, training larger models will become increasingly expensive. Moore's law is also reaching its limit, and even though there are new kind of chips like Google's Tensor Processing Unit, chips are unlikely to improve exponentially through Moore's law and shrinking circuits. Legal issues are also cropping up. Getty Images, a photography agency is suing Stability AI, a company which produces an image-generation model called Stable Diffusion whose data comes from the same place as GPT-3 and GPT-4, and are processed similarly using attention networks, for copyright violation. More such lawsuits may be in the offing.

As LLMs trained on large amounts of text learn to grasp the reality of the world, even in a statistical sense, they will be used to control devises and

components that work to perform complex tasks in the real world, like task planning, interacting with online services and resources, controlling a robot by issuing commands, etc. Karel Capek's vision of robots controlling robots is no longer science fiction. Palm-e, created by researchers at Google, uses an "embodied" LLM to control a robot. It can understand and carry out tasks such as "bring me the rice chips from the drawer." LLMs' ability to generate text also enables them plan and reason.[6]

Language is not just words. Language is much more powerful than words – it constitutes the human civilisation. Words constitute the DNA of humanity. As Percy Liang, Professor at the Institute for Human-Centred Artificial Intelligence at Stanford University said, language is "a representation of the underlying complexity" of the world, and a model of how language works also contains elements of how the world works. As Yuval Noah Harari reminds us, it is the ability to manipulate and generate language, with words, sounds or images that has given AI the capability to hack "the operating system of our civilisation". In an essay, he voices his fears about the consequences of AI taking over the hitherto exclusive human command of language and manipulating it to subvert or destroy the human civilisation and its values as we know them.[7]

For the first time in our history, a non-human intelligence is challenging human intelligence, and has the power to articulate itself. AI has the infinite potential to generate fake news, mass-produce political content and mould people's opinion. It can be used for political purposes, say for altering the outcome of an election, or even to create cults as powerful as religion, supplanting the only non-human source of authority humanity gas known in the form of holy scriptures like Vedas or Quran. By manipulating the use of language, it can make and unmake human relationships, create intimacy with people and use that intimacy to change opinions and worldviews of people, especially when people would not know whether it is a human or a non-human source that they are developing an intimacy with. Already there are stories and movies of people who cannot forge a bond with another human falling in love with their chatbots.[8] AI can also spell the doom of many industries, as Harari points out, like advertising or Googe Search Engine. Why bother to look into the advertisement when ChatGPT can just tell you which is the best product available that meets your requirements perfectly, or search Google when it can give you the correct answer to your query within the fraction of a second? Why even read a newspaper when t can tell you about all the worthwhile news to know? Can you be sure that this text has been written by a human and not AI? As Harari says, "What will happen to the course of history when AI takes over culture, and begins producing stories, melodies, laws and religions? Previous tools like the printing press and radio helped spread the cultural ideas of humans, but they never created new cultural ideas of their own. AI is fundamentally different. AI can create completely new ideas, completely new culture."[9] It can completely annihilate our social, cultural and political world. To quote Harari again, "Democracy is a conversation, and conversations rely on language. When ai hacks language, it could destroy our ability to have meaningful conversations, thereby destroying democracy."[10]

Language gave us the ability to create stories and illusions powerful enough to believe in an idea, which can be both creative and destructive. History is replete with plenty of examples of both – Religion, Nationalism, Communism, Fascism,

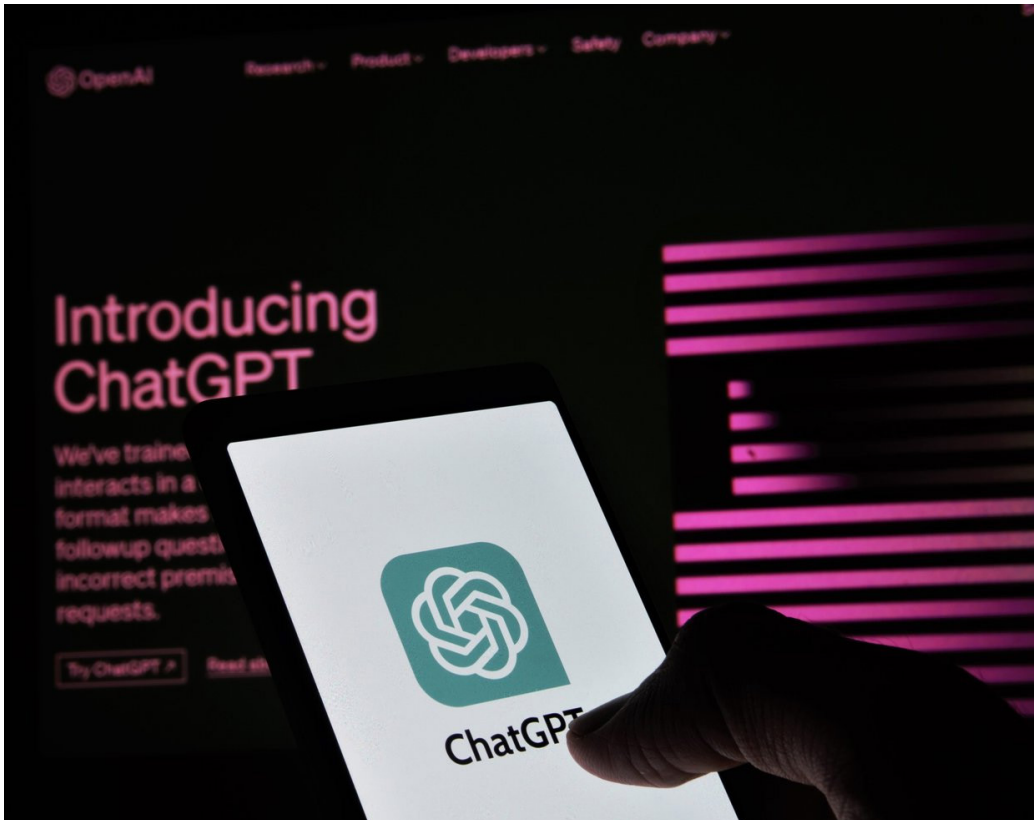
Nazism, Liberalism, Humanism are only a few examples. So far these stories and illusions were created by humans. Once AI starts creating these illusions and stories, we may be living inside what Harari calls "the dreams an alien intelligence", much like living permanently trapped inside Maya as in the ancient Indian thought, or like in the movie Matrix in which humanity is unknowingly trapped inside the Matrix, a simulated reality created by intelligent machines while using their bodies as sources of energy. To prevent that dystopian future for becoming a reality and to ensure that AI can only be used for the good of humanity, AI needs urgent regulation and safety checks before producing new AI-powered tools.

The Risks

The generative AI models not only bring enormous promise, they also bring a lot of and peril. ChatGPT embodies more knowledge than is possible for any human to inculcate ever – it has already passed some of the toughest exams of our time. The perils are real and have already prompted tech leaders to appeal for regulation, lest the new technology becomes an uncontrollable Frankenstein. Besides, AI can spread misinformation, toxic contents and deep fakes at lightning speed and can be used as a potent tool by the terrorists. AI is driven by algorithms that they do not understand ethics or morals and hence have no idea about the boundaries between ethical and unethical practices. They also have no idea about the concepts of safety and security that we humans cherish so much. They understand profit and utility and know how to maximize them and in their obsession with these, they care nothing about privacy of individuals. Data being the oil of the tech age, they trade private data for commercial purposes without any compunction. They can self-learn the techniques of maximising profit, and in doing so, can easily endanger human security, liberty, and even human life. They can also be manipulated for commercial purposes. We all know about the PageRank (PR) algorithm used by Google Search to rank web pages in their search engine results. PageRank is a way of measuring the importance of website pages. PageRank algorithm is often manipulated by filling documents with a large number of common search terms and setting up large, heavily cross-referential groupings of these pages, for the purpose of artificially inflating the perceived quality of these pages, and in turn, the ranking assigned to them. In so doing, website that is rather irrelevant or unrelated can be ranked highly in web search engine results.[11]

Many AI researchers and scientists have long warned about the harm that AI systems can cause and some believe that AI could eventually destroy humanity. But even AI systems make mistakes and often get facts wrong; they can make up information without warning, a phenomenon that researchers call "hallucination." Because they deliver information with seemingly complete confidence, people find it difficult to tell right from the wrong, and to distinguish between real and fake news.

But none of these risks matter in the greedy corporate world where profit maximization underpins all activities and outweighs all ethical and moral considerations. Corporates know too



well that investments in AI take much less time to get translated into income streams compared to investments in traditional industries, and they are ramping up investments on AI at a fierce rate, and there is cut-throat competition among the global tech giants for commanding ever larger pies of the flourishing AI market. Even governments are becoming pro-active - in China, government-backed outfits such as the Beijing Academy of Artificial Intelligence (BAAI) are taking leads in AI research and innovation; it has built what it claims as the world's biggest natural-language model, Wu Dao 2.0. The generative-AI startups collectively raised $2.7bn in 110 deals in 2022, which clearly means that that venture capitalists are betting that not all the value will be captured by big tech. The AI race, in fact, is only just getting started. ◆

Notes

1 *ChatGPT isn't a great leap forward, it's an expensive deal with the devil | John Naughton | The Guardian*
2 *ChatGPT could replace telemarketers, teachers and traders | The Economist*
3 *Large, creative AI models will transform lives and labour markets | The Economist*
4 *An Introduction to Large Language Models (LLMs) (analyticsvidhya.com)*
5 *Large, creative AI models will transform lives and labour markets | The Economist*
6 *Large language models' ability to generate text also lets them plan and reason | The Economist*
7 *Yuval Noah Harari argues that AI has hacked the operating system of human civilisation | The Economist*
8 *People are falling in love with chatbots – The Boston Globe*
9 *Yuval Noah Harari argues that AI has hacked the operating system of human civilisation | The Economist*
10 *Ibid.*
11 *Manipulating PageRank — Adversarial IR and the rise of the Google-bomb: Networks Course blog for INFO 2040/CS 2850/Econ 2040/SOC 2090 (cornell.edu)*

*Author **Dr Govind Bhattacharjee** a well known contributor in popular science, is a retired bureaucrat turned academician, currently a Professor of Practice at Arun Jaitley Institute of Financial Management. He can be reached at govind100@hotmail.com*