The conundrum of ChatGPT

Subhashis Banerjee *

March 16, 2023

ChatGPT launched by OpenAI in November 22 is the closest Artificial Intelligence (AI) has come to passing the Turing test of Intelligence, on whether a machine can engage in a conversation with a human without being detected as a machine. Its capabilities are multifaceted. It can write essays and letters on specified topics, engage in coherent conversations with humans on most subjects, write snippets of computer programs or poetry in specified styles, and seemingly even do logical reasoning and make new hypotheses.

More such tools with more enhanced capabilities will inevitably follow. While whether such machines can really do human-like reasoning or truly exhibit intelligence are open questions in computer science and philosophy, it is undeniable that such tools will have a profound impact on education and skill building as we know them today. They not only provide new opportunities but also new challenges. Without doubt, the society needs to understand, debate and evolve methods to deal with such disruptions.

It is but natural for a bewildered society to wonder that if indeed ChatGPT can write code and coherent essays on most topics, then what values remain in learning these skills? Do we then run the risk that many jobs – especially some of the more routine ones – will be replaced by such large language models? How then should our education processes be transformed, or how may the use of such tools be regulated?

These indeed are difficult questions, and it may be helpful to first try to understand at a broad level how such large language models work. GPT-3, which ChatGPT is based on, was trained on over 45 Terabytes of text data to learn the patterns and relationships between various parts of text and their context. This learning is represented in terms of over 175 billion internal parameters – or 'weights' – of the 'transformer' model. Then, given a new prompt or a question, the model calculates the probability of the next word it should output based on these internal weights. Context flows in through a special attention mechanism in the transformer.

It is amazing that such models work so well, and most of us would not have believed it possible a decade back. However, the models are purely statistical in nature, and it is not clear that capabilities of deductive reasoning can ever be acquired through generalisations from a purely inductive process of learning from examples. Indeed, mistakes – even factual ones – are common, and it is not difficult to make such models fail on deductive tasks.

Neither can such models do abductive reasoning. That is, given an observation the model cannot truly generate causal hypotheses. It is widely believed that such hypotheses generation in a true sense is not possible purely from generalisations from correlations and associations. Also, it is well known that causal inference is impossible from such data correlations without counterfactual reasoning or conducting experiments.

*Professor of Computer Science, Ashoka University and Professor of Computer Science and Engineering, IIT Delhi. Email: suban@ashoka.edu.in

Moreover, intelligence cannot possibly be defined purely behaviourally, and, even if such language models acquired these capabilities, they would still be epistemologically incomplete for human reasoning, which involves complex cognitive processes. Capabilities of critical thinking and ethical reasoning cannot simply be acquired from a machine that merely learns patterns and associations. Thus, over-reliance on such models is bound to be severely limiting for both students and teachers.

However – if used well – such language models may actually help us to focus more on reasoning instead of rote learning of facts, and to separate the trite from the original. Students may use them as tools that make learning more efficient, and teachers may calibrate their teaching to concentrate more on critical reasoning than facts.

The other ethical conundrum that emerges is that who really is responsible for the answers provided by ChatGPT like models, and who should be accountable for the opinions gathered and the mistakes made? Should it be OpenAI and the likes? Or are the opinions truly democratised wisdom curated from millions of written text that were used for training the language models? If so, who all are responsible for these opinions and how they may be reviewed? Or is it the case that ultimately everything is responsibility of just the user?

ChatGPT is a great technological innovation, but there are some real dangers of innocent and ordinary readers being seduced by the surface coherence and sophistication of the text generated, and accepting things on their face value without closer examination. One only wishes that some responsible and ethical directions also originate from the BigTech companies when they unleash such products, rather than leave it to some hapless students and a bewildered society to deal with them. After all, artifacts rarely are politics-free, and to suggest that they simply are tools and it is upto the users to use them well appears to be simplistic and deeply problematic.