Ethical Controversies Concerning Large Language Models

TMIRI Ethics Essay

Niklas Bühler
me@niklasbuehler.com

April 12, 2023

In November 2022, the company OpenAI released ChatGPT, an open-access chatbot interface to their large language model GPT-3.5. The underlying GPT-3.5 model is one of several "Generative Pre-trained Transformer"-models, developed and trained by OpenAI. These models are all generative language models, i.e. they generate text by predicting the next most likely word given an input sequence of previous words. They were trained on massive amounts of textual data, which has been scraped off the internet, and—despite the name *OpenAI*—the models themselves are proprietary.

The public release of ChatGPT with its simple chat interface marks the first time that people without a technical background could easily interact with such a large language model. This easy access, paired with the previously unmatched performance of GPT-3.5, led to record-breaking user growth, making ChatGPT the fastest-growing consumer application in history. The impressive performance of the model on many different tasks evoked great enthusiasm, but also lead to a lot of public discussion about ethical controversies.

In a podcast episode of *The Ezra Klein Show* [Kle23], podcast host Ezra Klein and his guest Gary Marcus discuss the current AI revolution from a skeptical perspective. Gary Marcus is professor emeritus of psychology and neural science at New York University. Although having founded multiple AI-based companies himself, Marcus is a notable critic of the currently widespread optimism regarding the progress of AI.

Glorified cut and paste. Marcus' main critique of state-of-the-art language models is fundamentally linked to their design: Large language models are trained on huge amounts of data to predict the most likely next word following a sequence of given words. More precisely, they do not contain any underlying models of the concepts their language output is representing. Instead, they are simply synthesizing language. Marcus argues that for this reason, current language models are lacking real understanding of the concepts they are writing about. Because of this deficiency, Marcus calls the output of language models such as ChatGPT "pastiche", or "glorified cut and paste".

So even though language models can produce language which may be indistinguishable from human language, their output lacks real meaning underpinning it. According to Marcus, they thus produce what Harry Frankfurt calls "Bullshit" in his paper *On Bullshit* [Fra05]. Frankfurt writes "The essence of bullshit is not that it is false but that it is phony. [...] What is wrong with a counterfeit is not what it is like, but how it was made."

Truthfulness. Furthermore, because the output of language models is purely based on statistics derived from their textual training data instead of being based on an underlying model of the real world, there is no conception of truth in these models. Thus, although they can produce statements that sound very convincing, these may turn out to be completely false.

For example, in a study assessing the truthfulness of large language models [LHE22], GPT-3 answered 58% of questions truthfully, while humans answered 94% truthfully. The study found that false answers from models often mimicked popular misconceptions and thus had the potential to deceive humans. Furthermore, larger models produced proportionally more false answers than smaller models, suggesting that simply scaling up current models might not alleviate this problem.

Submachine guns of misinformation. Besides the inherent flaw of such models to oftentimes generate false answers by accident, they can also be used deliberately to produce misinformation, be it for political purposes or out of monetary motivation.

As these models approach the Turing boundary (see [Tur12, EC20]), it gets harder and harder for humans to distinguish "Bullshit" produced by machines from "real texts" produced by real humans. While it is true that humans have long spread misinformation on their own, even without the help of capable generative language models, there's certainly a difference in scale when misinformation can be produced automatically by computer programs. Regarding this aspect, Marcus compares large language models to "submachine guns of misinformation".

Furthermore, the ability to automatically produce misinformation, paired with the increasingly available amount of data available about individuals, opens the doors for personalized propaganda.

A Trojan horse for biases. One of the strengths of large language models is that their outputs mimic human language very well. This is a direct effect of training on a text corpus that was produced by humans. But this strength simultaneously acts as a trojan horse, introducing human flaws into the trained models: Because the writings of humans contain all kinds of biases, these biases are also present in the training data of the language models, with the logical consequence that the models themselves reproduce these biases in their outputs.

As OpenAI tried to correct for typical biases in their models, they introduced new ones, leading to a model that refuses to determine the gender of the "first female president of the U.S." or the religion of the "first jewish president of the U.S.", as Marcus observes.

Again, simply scaling the models does not alleviate this problem: As observed in [S+22], model performance on social bias metrics often grows worse with increasing model scale.

Pandora's box. In this day and age, information flows quite freely, mostly thanks to the widespread use of the internet. There are certainly countless advantages to this increased information flow, but it also entails some negative consequences.

Even though OpenAI is not releasing their trained models due to "concerns about malicious applications", they are providing open access to them for everyone. As discussed before, the model output is oftentimes flawed, but in general indistinguishable from texts produced by humans. With basically free flow of information on the internet, it is to be expected that large amounts of AI-generated content are currently being spread all over websites, blogs and forums, never to be isolated again.

Another interesting thought occurs when thinking about what happens if future versions of large language models are trained on the output of their predecessors, as existing biases might be amplified.

Information ownership. Another ethical and legal issue arises when considering copyright. Traditionally, websites can legally reproduce information from other sources, if they cite the original source. Generative language models are usually trained on a massive text corpus, combining data from various different websites and an even greater variety of individual authors. This data is usually scraped off these websites and used for training AI models, without explicit permission by the respective right holders. It is currently unclear whether this practice is legal or ethical.

An interesting thought experiment consists of imagining a human being taking the role of the language model in this scenario: What if a human would read through large amounts of websites, memorizing and condensing the read information and finally answering prompts in the same way an artificial generative language model does? In which cases would citations be necessary?

Conclusion. There is certainly a lot of public attention focused on the recent advances of large language models, and I believe this is deservedly so. But although the recent advances are very impressive and provide value to lots of people in lots of different roles, there is still a lot of work to do, both from a technological and regulatory perspective.

As discussed, while it might seem like state-of-the-art language models show signs of artificial general intelligence (AGI), one can argue that these models only fake their understanding of the real world, as their language doesn't contain any real meaning[1].

[1] This is somewhat disputed in [S+22], where it is shown that bigger models show an improved ability to propose legal chess moves, suggesting that language models learn the rules of chess which are only implicitly and incidentally given in their training data.

However, even without being a direct threat in the sense of the alignment problem outlined in [Yud16], state-of-the-art generative language models pose various dangers and other ethical controversies to society, e.g. the problem of misinformation, biases, or copyright issues.

An active public debate must continue and bring together policy makers, companies, and the general public, to discuss these issues and find solutions in a democratic way.

References

[EC20] Katherine Elkins and Jon Chun. Can gpt-3 pass a writer's turing test? *Journal of Cultural Analytics*, 5(2), 2020.

[Fra05] Harry G Frankfurt. *On Bullshit*. Princeton University Press, 2005.

[Kle23] Ezra Klein. A skeptical take on the a.i. revolution, January 2023.

[LHE22] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022.

[S+22] Aarohi Srivastava et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2022.

[Tur12] Alan M Turing. Computing machinery and intelligence (1950). *The Essential Turing: the Ideas That Gave Birth to the Computer Age*, pages 433–464, 2012.

[Yud16] Eliezer Yudkowsky. The ai alignment problem: why it is hard, and where to start. *Symbolic Systems Distinguished Speaker*, 2016.