Is AI Undermining Trust Online? ChatGPT, Large Language Models, and Disinformation

The aim of Disinfo Radar's Rapid Response Briefs is to identify new and noteworthy disinformation technologies, tactics and narratives. Such cases may include the identification of a new technology that may have harmful uses.

Background

The AI chatbot system ChatGPT has been flooding social media in recent weeks, having been used more than 1 million times in a span of only 5 days. The text generator, bearing resemblance to search engines, can provide answers to a user's burning questions in split seconds, but brings with it the potential to impact not only our general online experience but also corporate workflows, the future of academia and the use of Google Search.

What if ChatGPT mixes correct information with false information, or creates completely made-up disinformation? How will users know when text is authentic or stems from an advanced language model? Is this the end of Google Search as we know it? Not quite — but it can pass the Turing test[1], opening doors for inaccurate information.

What is ChatGPT?

ChatGPT is a natural language generation model created by OpenAI and trained to follow instructions from a text prompt. The special feature of ChatGPT is its dialogue format: It is possible to communicate via text input with the language model, as if one were talking to a human. The dialogue format enables ChatGPT to "answer follow-up questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests," according to the OpenAI website.

The machine learning model is trained using huge amounts of text and Reinforcement Learning from Human Feedback (RLHF). How does this work? Through training, the

1 The Turing test was proposed by computer scientist Alan Turin in 1950 to gauge intelligence: Can a human being talking to another human and a computer system tell which one is which?

model learns which meaning is associated with certain questions or demands. The bot is then optimised with the help of human trainers, who evaluate its answers and, thus, through positive feedback, gradually teach the model which formulations are desirable.
There are several different large language models (LLM) that analyse and are trained on large amounts of text, including ChatGPT's foundation, Generative Pre-Trained Transformer 3 (GPT-3), which can imitate human-generated text, and the model landscape is constantly growing, with Google's recent LaMDA and Open AI's GPT-4, which is to be released in 2023. ChatGPT expands beyond already existing models, due to its communicative chat function.

Why Can This Technology Be Misused for Disinformation Purposes?

While this technology has not yet (that we know of) found its way into global disinformation campaigns, large language models have the potential to become more important to disinformation efforts. Research by Georgetown University, in Washington, D.C., has already illustrated how difficult it is for people to distinguish genuinely written texts from GPT3-generated ones – and how the latter can be used to generate disinformation. As such models continue to improve and it becomes possible to easily and in a very short time create fake content that is difficult to tell apart from authentic content, people could use them to generate and spread mis- and disinformation. One plausible scenario would be bots contributing misleading text to the comment sections of posts.

Whereas ChatGPT's responses might seem useful and authoritative, they can be entirely inaccurate, based on flawed input data, or can blend fact with fiction. Large language models are prone to taking what they have learnt and reshaping it into something new – regardless of its accuracy. ChatGPT could provide a plausible answer to one question, but nonsense to another formulated slightly differently. Further, the same input prompt can generate two different responses. If people take the response created by the model at face value, without verifying information from original sources, this can threaten evidence-based online discourse. An innocent user can easily be fooled in believing the information provided by ChatGPT is entirely reliable.

A language model such as GPT-3 is primarily intended to provide answers that are as plausible as possible in the context of conversations. Instead of simply replying „I don't know the answer", the algorithm invents answers and confidently presents them as facts.

While there have been efforts to mitigate the spread of inaccurate information by restricting the input data of the model, the rise of generative AI[2] will make authenticity online even harder to determine than it already is. Meta recently removed an online preview of its own large language model, Galactica, because it repeatedly generated incorrect and biased information. Whereas companies like Google, Meta and OpenAI can expedite the development of this technology,

2 Un- or semi-supervised machine learning for creating new content

they cannot prevent open-source replica models that do not employ any moderation or self-regulation from emerging and spreading even more misinformation.

How Can Generative AI Merge Synthetic Creation Processes?

It is likely, therefore, that we are entering a phase where there will be a flood of texts that present false information. With generative AI such as ChatGPT, the merging of synthetic creation processes is far more plausible, rendering matters worse if fully automatised disinformation pipelines were to be established. This would allow disinformation actors to produce more coherent fake evidence that will be increasingly hard to debunk more quickly. Other advantages are the fluency, richness and speed of production that low-level actors struggle to replicate. Another possibility would be the combination of text created by large language models with text-to-image generators, such as DALL-E 2, to quickly create convincing disinformation campaigns. Such an automatisation of the process can replace the tedious creation of false content and increase the complexity of disinformation strategies in the online environment.

Expected Threats:

- Mis- and disinformation produced and convincingly presented by large language models can spread when taking responses provided by machine learning at face value.

- Text generation models can change the speed and richness of the production of textual disinformation.

- Without moderation, a replica large language model chatbot that is consistently trained with false statements can subsequently produce such false statements in dialogue.

- The automatisation of synthetic creation processes can produce coherent disinformation campaigns more quickly.

How to Respond?

- AI service providers should restrict the use of their large language models for harmful content. Like every product from OpenAI, ChatGPT has security mechanisms for output data in place, using moderation to block inappropriate content. These, however, can be easily circumvented with creative questions, so that they still respond to harmful text prompts. Regular and standardised model evaluation and risk assessment to identify potential harms should be introduced. This would ensure a more responsible and sustainable use of the model.

- Moreover, the large-scale input data that feeds the algorithms is unfiltered, thus often confirming bias and stereotypes. These flaws will be passed on to subsequent models. AI service providers should, as suggested by OpenAI, enable the provision of user feedback on problematic model outputs to help both the industry and regulators to standardise reporting mechanisms. Filtered training data could then help mitigate the production of harmful content.

- AI service providers and algorithm developers should build large language models that can accurately determine when it is confident about an answer the model provides and when it is not, so as to reduce the risk of spreading misinformation. Providing a statistical confidence interval as an indicator of the content's accuracy would be one option.

Date: December 2022

This Rapid Response Brief was written by Lena-Maria Böswald, Digital Democracy Programme Officer, and is part of Democracy Reporting International's Disinfo Radar project, funded by the German Federal Foreign Office. Its contents do not necessarily represent the position of the German Federal Foreign Office.

About Democracy Reporting Inte rnational

DRI is an independent organisation dedicated to promoting democracy worldwide. We believe that people are active participants in public life, not subjects of their governments. Our work centres on analysis, reporting and capacity-building. For this, we are guided by the democratic and human rights obligations enshrined in international law. Headquartered in Berlin, DRI has offices in Lebanon, Libya, Myanmar, Pakistan, Sri Lanka, Tunisia, and Ukraine.