Published on *The National Law Review* https://natlawreview.com

ChatGPT ‘Hallucinates’ and Other Conclusions from OpenAI’s Paper on Safety Concerns

Article By:

Malcolm Dowden

ChatGPT is a powerful tool for work and study purposes. From an enterprise perspective, being able to find answers to complex questions by simply typing in a prompt is attractive. However, it is essential to consider the veracity of the information obtained. A recent paper published by OpenAI – the developer of ChatGPT – explores the various safety, privacy and cybersecurity concerns that their AI tool creates, as well as the actions they have taken to mitigate the potential harms. Although some solutions are proposed, many challenges remain, highlighting the need to ensure that any business deployment of ChatGPT or other generative AI services comes with adequate controls over data preparation, prompts and the screening of ChatGPT’s responses.

Hallucinations

In the words of OpenAI, ChatGPT has the tendency to “hallucinate”. This is defined as producing content that is nonsensical or untruthful in relation to certain sources. For example, when requested to make a summary of an article, ChatGPT produced – within the correct information about its content – some made-up paragraphs with information that did not appear in the original source.

The implications of these hallucinations are widespread. When asked to provide a legal argument, ChatGPT provided complete citations to case law that did not exist. Even when given specific documents to draw information from (for example, an insurance policy), the appearance of hallucinations in the answer could not be ruled out.

Counterintuitively, the risk of hallucinations going unchallenged increases the more ChatGPT is used. OpenAI rightly identifies that, as the model develops and becomes more convincing, users will start to place more trust on the information it tells them. This opens the possibility of hallucinated “facts”, including wholly fabricated legal precedents, being treated as truths.

Overreliance

Linked to the risks identified above, is the issue of overreliance. As it develops, ChatGPT becomes ever more believable by giving users complex and detailed answers, gaining authority in the process. This is especially the case when the user is researching an area that is not within their expertise, making it more difficult to verify the information provided by the chat.

OpenAI also warns that the model's tendency to hedge in its responses, appearing cautious in its approach, can also increase the user's trust and reliance on its answers. When ChatGPT seems 'careful', we're more likely to believe it (and its hallucinations).

OpenAI has attempted to tackle this problem by making ChatGPT more rigorous when rejecting requests that violate its policies. However, the paper also shifts responsibility to developers using their tools, who are advised to issue guidance on how to use the model and its limitations. The message here is that it is the user's responsibility to be critical of the information provided by ChatGPT and distinguish fact from fiction – which considering the complexity and detail of some of the model's hallucinations, can become a complicated task.

Harmful content

OpenAI identified that ChatGPT was able to generate harmful content which can have an impact in the real world, exploiting and harming individuals. This includes:

- Hate speech and harassing, demeaning and hateful content;
- Information on how to plan attacks or acts of violence, including how to find illegal content; or
- Graphic material.

Even though OpenAI has worked to improve its systems refusals, they state that intentional probing of the model can lead to these results. OpenAI can work to block the prompts or questions they anticipate, but they can't guess all possible ways of asking.

They recognise that the increase in the model's capabilities, and its widespread implementation, will mean that the challenges identified are not just likely but imminent. OpenAI concludes by encouraging further research into AI literacy, economic and social resilience and participatory governance, to ensure a smoother transition into the use of these models.

While the actions recommended are necessary, the development and implementation of these measures is likely to be outpaced by the ever-increasing use and normalisation of these AI tools. How many of those who have used ChatGPT knew that it could 'hallucinate'? These are important points to consider when deciding whether to allow the use of the tool in your workplace or trust it for your next study project.

ChatGPT may look like it has all the answers, but as the CEO of OpenAI noted, this is so far "a misleading impression of greatness".

Considerations for the future

In a new episode of our Privacy Law Podcast, legal technologist Jonathan Bowker explains the data preparation, controls and risk mitigation measures that should form an essential part of any enterprise deployment of generative AI. Listen to the Podcast here.

National Law Review, Volume XIII, Number 178

Source URL:https://natlawreview.com/article/chatgpt-hallucinates-and-other-conclusions-openai-s-paper-safety-concerns