Hallucinating chatbots

February 20, 2023

In news– It has been warned by google that artificial intelligence in chatbots can sometimes lead to "hallucination".

About hallucinating chatbots-

- **Hallucination in AI chatbots is when a machine provides convincing but completely made-up answers.**
- It is not a new phenomenon and developers have warned of AI models being convinced of completely untrue facts, responding to queries with madeup answers.
- In 2022, Meta released their AI conversational chatbot called BlenderBot 3.
- At the time the company shared that **BlenderBot 3** was capable of searching the internet to chat with users about any topic and would learn to improve its skills and safety through feedback from users.
- However, even at that time, Meta engineers had warned that the **chatbot should not be relied upon for factual information and that the bot could apparently "hallucinate".**
- An example of this was seen in 2016 when after being live on Twitter for just 24 hours, **Microsoft's chatbot Tay** started parroting racist and misogynistic slurs back at users.
- The chatbot, designed as an experiment in "conversational understanding", could be manipulated by users by just asking it to "repeat after me".

Why do AI chatbots start hallucinating?

- **A defining feature of sophisticated generative natural language processing (NLP) models, hallucinations, can occur because these models require the capability to rephrase, summarise** and present intricate tracts of text

without constraints.

- This raises the **problem of facts not being sacred** and they can be treated in contextual form when sifting through information**. An AI chatbot could possibly take widely available information rather than factual information as an input**. The problem becomes especially acute when **complex grammar or arcane source material is used.**
- Therefore, AI models can start presenting and even believing in ideas or information that may be incorrect but which are fed to them by a large number of user inputs. And since **these models are unable to distinguish between contextual information and facts,** they respond to queries with incorrect answers.
- For example, when asked “What does Albert Einstein say about black holes?” AI models can return a quote made famous on the Internet rather than factual information based on Einstein’s research.

Further reading: https://journalsofindia.com/chatgpt-vs-googles-bard/